UNIVERSIDAD DE LA RIOJA

# TRABAJO FIN DE ESTUDIOS

Título

## Modelos probabilísticos de predicción eléctrica a corto plazo en una planta fotovoltaica


Autor/es

OMAR RADA GARCÍA

Director/es

LUIS ALFREDO FERNÁNDEZ JIMÉNEZ y EDUARDO GARCÍA GARRIDO ,

Facultad

Escuela Técnica Superior de Ingeniería Industrial


Titulación

Grado en Ingeniería Eléctrica


Departamento

INGENIERÍA ELÉCTRICA


Curso académico

2018-19





publicaciones.unirioja.es
E-mail: publicaciones@unirioja.es

# PROYECTO FINAL DE GRADO EN INGENIERÍA ELÉCTRICA

## MODELOS PROBABILÍSTICOS DE PREDICCIÓN ELÉCTRICA A CORTO PLAZO EN UNA PLANTA FOTOVOLTAICA

OMAR RADA GARCÍA

UNIVERSIDAD
DE LA RIOJA

# Índice de Contenidos

*1. Resumen* ..... *6*

*2. Abstract* ..... *7*

*3. Introducción* ..... *8*

3.1 Objetivos del trabajo ..... 9

3.2 Estructura del trabajo ..... 10

*4. Situación actual de las energías renovables* ..... *10*

4.1 Situación de la energía solar fotovoltaica en España ..... 11

*5. Necesidad de la predicción de la energía solar* ..... *13*

*6. Horizontes de predicción* ..... *15*

*7. Herramientas utilizadas para el desarrollo de los modelos de predicción* ..... *16*

7.1 Programa Microsoft Excel ..... 16

7.2 Programa R ..... 16

7.2.1 La interfaz Rstudio ..... 17

*8. Desarrollo del trabajo* ..... *19*

8.1 Planta fotovoltaica de Alcolea del Río ..... 19

8.2 Realización de la muestra ..... 20

8.2.1 Predicciones meteorológicas ..... 21

8.2.2 Potencias producidas por la planta fotovoltaica ..... 23

*9. Criterios de evaluación de los modelos de predicción* ..... *26*

9.1 Modelos determinísticos o puntuales ..... 27

9.2 Modelos de predicción probabilísticos ..... 28

*10. Modelos de predicción* ..... *31*

10.1 Modelo climatológico ..... 31

10.1.1 Código para el análisis mediante la herramienta Rstudio ..... 36

10.2 Modelo persistente probabilístico ..... 41

10.2.1 Modelo día similar (hoy – hoy) ..... 41

10.2.2 Modelo día similar (hoy – mañana) ..... 49

10.3 Regresión de Cuantiles ..... 57

10.3.1 Ventajas de la regresión cuantílica ..... 59

10.3.2 Código para el análisis mediante la herramienta Rstudio ..... 60

10.4 Regresión Lineal Múltiple ..... 64

10.4.1 Código para el análisis mediante la herramienta Rstudio ..... 71

10.5 Random Forest (Bosques aleatorios) ..... 73

**10.5.1 Árboles de clasificación** ..... 73
**10.5.2 Bagging y boosting** ..... 73
**10.5.3 Concepto Random Forest** ..... 74
**10.5.4 Validación cruzada (Cross validation)** ..... 81
**10.5.5 Código para el análisis mediante la herramienta Rstudio** ..... 84

***11. Comparación de los modelos de predicción probabilístos*** ..... ***88***

***12. Conclusiones*** ..... ***92***

***13. Bibliografía y webgrafía*** ..... ***94***

**13.1 Listado de la bibliografía** ..... **94**

**13.2 Listado de la webgrafía** ..... **95**

# Índice de Imágenes

ILUSTRACIÓN 1: POTENCIA SOLAR FOTOVOLTAICA INSTALADA ACUMULADA EN ESPAÑA ..... 12
ILUSTRACIÓN 2: POTENCIA SOLAR FOTOVOLTAICA INSTALADA ANUALMENTE EN TODO EL MUNDO ENTRE LOS AÑOS 2000 Y 2017 (EXPRESADA EN GW) ..... 13
ILUSTRACIÓN 3: MUESTRA ENTORNO DE TRABAJO DE RSTUDIO ..... 17
ILUSTRACIÓN 4: REPRESENTACIÓN DESDE VISTA AÉREA DE LA DISTRIBUCIÓN DE LOS MÓDULOS FOTOVOLTAICOS ..... 20
ILUSTRACIÓN 5: REPRESENTACIÓN REAL DE LOS MÓDULOS FOTOVOLTAICOS DE ALCOLEA DEL RÍO .. 20
ILUSTRACIÓN 6: DOMINIOS ANIDADOS QUE DELIMITAN SOBRE QUE ÁREA SE REALIZARÁN LAS PREDICCIONES METEOROLÓGICAS ..... 22
ILUSTRACIÓN 7: SITUACIÓN DE LOS 4 PUNTOS MÁS CERCANOS A ALCOLEA DEL RÍO PARA HACER LA INTERPOLACIÓN CUADRÁTICA DE LA PREDICCIÓN METEOROLÓGICA ..... 23
ILUSTRACIÓN 8: REPRESENTACIÓN DE LA MUESTRA CORRESPONDIENTE AL GRUPO DE ENTRENAMIENTO CON "OUTLIERS" ..... 24
ILUSTRACIÓN 9: REPRESENTACIÓN DE LA MUESTRA CORRESPONDIENTE AL GRUPO DE TEST ..... 25
ILUSTRACIÓN 10: DISTRIBUCIÓN PROBABILÍSTICA DEL MODELO CLIMATOLÓGICO ..... 34
ILUSTRACIÓN 11: REPRESENTACIÓN DE TRES DÍAS CONSECUTIVOS DEL MODELO CLIMATOLÓGICO FRENTE A LA PRODUCCIÓN REAL ..... 36
ILUSTRACIÓN 12: HISTOGRAMA PERTENECIENTE AL GRUPO DE ENTRENAMIENTO, MODELO CLIMATOLÓGICO ..... 39
ILUSTRACIÓN 13: HISTOGRAMA PERTENECIENTE AL GRUPO DE TEST, MODELO CLIMATOLÓGICO ..... 40
ILUSTRACIÓN 14: HISTOGRAMA GRUPO DE TEST, MODELO PERSISTENTE PROBABILÍSTICO "HOY-HOY" ..... 49
ILUSTRACIÓN 15: HISTOGRAMA GRUPO DE TEST, MODELO PERSISTENTE PROBABILÍSTICO "HOY-MAÑANA" ..... 56
ILUSTRACIÓN 16: EJEMPLO DE LA RESOLUCIÓN MEDIANTE LA HERRAMIENTA SOLVER ..... 66
ILUSTRACIÓN 17:REPRESENTACIÓN DE LA POTENCIA DEL GRUPO DE ENTRENAMIENTO FRENTE A LA POTENCIA DEL MODELO, MODELO REGRESIÓN LINEAL MÚLTIPLE ..... 67
ILUSTRACIÓN 18: COMPARATIVA ENTRE LOS PERCENTILES OBTENIDOS FRENTE A LA POTENCIA REAL PRODUCIDA, MODELO REGRESIÓN LINEAL MÚLTIPLE ..... 70
ILUSTRACIÓN 19: AJUSTE LÍNEA DE REGRESIÓN ENTRE LOS VALORES REALES Y LOS PREDICHOS ..... 70
ILUSTRACIÓN 20: COMPARATIVA DE LA POTENCIA REAL PRODUCIDA FRENTE A LA PREVISIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE ..... 71
ILUSTRACIÓN 21: EJEMPLO DEL FUNCIONAMIENTO DEL RANDOM FOREST ..... 75
ILUSTRACIÓN 22: REPRESENTACIÓN DEL CRPS DEL GRUPO DE ENTRENAMIENTO FRENTE AL CRPS DEL GRUPO DE TEST, MODELO RANDOM FOREST ..... 79
ILUSTRACIÓN 23: REPRESENTACIÓN DEL MAE DEL GRUPO DE ENTRENAMIENTO FRENTE AL MAE DEL GRUPO DE TEST, MODELO RANDOM FOREST ..... 79
ILUSTRACIÓN 24: REPRESENTACIÓN DEL $R^2$ DEL GRUPO DE ENTRENAMIENTO FRENTE AL $R^2$ DEL GRUPO DE TEST, MODELO RANDOM FOREST ..... 80
ILUSTRACIÓN 25: PROCESO A REALIZAR PARA ELABORAR EL MODELO MEDIANTE K-FOLD CROSS VALIDATION ..... 83
ILUSTRACIÓN 26: REPRESENTACIÓN DE LOS ERRORES PARA LOS MODELOS DE PREDICCIÓN DEL DÍA D ..... 89
ILUSTRACIÓN 27: REPRESENTACIÓN DE LOS ERRORES DE LOS MODELOS DE PREDICCIÓN PARA EL DÍA D+1 ..... 91

# Índice de Tablas

TABLA 1: EJEMPLO DE LAS MUESTRAS UTILIZADAS PARA EL DESARROLLO DE LOS MODELOS DE PREDICCIÓN ..... 26
*TABLA 2: EJEMPLO DE COMO SE HAN CALCULADO LOS PERCENTILES PARA DIFERENTES HORAS DEL DÍA* ..... 33
TABLA 3: REPRESENTACIÓN DE LOS PERCENTILES DE TRES DÍAS CONSECUTIVOS JUNTO CON LA PRODUCCIÓN REAL, MODELO CLIMATOLÓGICO ..... 36
TABLA 4: RESULTADOS OBTENIDOS PARA LOS GRUPOS DE ENTRENAMIENTO Y TEST, DESPUÉS DE EJECUTAR RSTUDIO ..... 38
TABLA 5: ELECCIÓN VENTANA FIJA ÓPTIMA, MODELO PERSISTENTE PROBABILÍSTICO "HOY-HOY" ..... 45
*TABLA 6: RESULTADOS DE LOS GRUPOS DE ENTRENAMIENTO Y TEST, MODELO PERSISTENTE PROBABILÍSTICO "HOY-HOY"* ..... 48
TABLA 7: ELECCIÓN VENTANA FIJA ÓPTIMA, MODELO PERSISTENTE PROBABILÍSTICO "HOY-MAÑANA" ..... 52
*TABLA 8: RESULTADOS GRUPO DE ENTRENAMIENTO Y TEST, MODELO PROBABILÍSTICO PERSISTENTE "HOY-MAÑANA"* ..... 55
TABLA 9: RESULTADOS GRUPO DE ENTRENAMIENTO Y DE TEST, MODELO REGRESIÓN DE CUANTILES 63
TABLA 10: PARÁMETROS PARA DETERMINAR LA ECUACIÓN DE CÁLCULO DE LA POTENCIA PRODUCIDA PREVISTA ..... 66
TABLA 11: EJEMPLO DE PERCENTILES CALCULADOS MEDIANTE HOJA DE CÁLCULO EXCEL, MODELO REGRESIÓN LINEAL MÚLTIPLE ..... 69
*TABLA 12: RESULTADOS GRUPO DE ENTRENAMIENTO Y DE TEST, MODELO REGRESIÓN LINEAL MÚLTIPLE* ..... 72
TABLA 13: EJEMPLO DE LOS DATOS OBTENIDOS MEDIANTE RSTUDIO EN EL PRIMER INTENTO, MODELO RANDOM FOREST ..... 78
TABLA 14: RESULTADOS OBTENIDOS DE LOS 3-FOLDS A TRAVÉS DE RSTUDIO ..... 86
TABLA 15: VALORES MEDIOS OBTENIDOS DE LOS 3-FOLDS, DETERMINACIÓN DE LA COMBINACIÓN DE NODOS Y VARIABLES ÓPTIMA ..... 87
*TABLA 16: RESULTADOS FINALES DE PREDICCIÓN DEL GRUPO DE ENTRENAMIENTO Y TEST, MODELO RANDOM FOREST* ..... 88
*TABLA 17: COMPARACIÓN DE LOS MODELOS PARA EL DÍA D (PREDICCIONES PARA HOY)* ..... 89
TABLA 18: COMPARACIÓN DE LOS RESULTADOS FINALES ENTRE LOS MODELOS DE PREDICCIÓN PARA EL DÍA D+1 ..... 91

# Índice de Ecuaciones

ECUACIÓN 1: FÓRMULA QUE REPRESENTA EL MODO DE CALCULAR EL MAE ..... 27
ECUACIÓN 2: FÓRMULA QUE REPRESENTA EL MODO DE CALCULAR EL RMSE ..... 28
ECUACIÓN 3: FÓRMULA QUE REPRESENTA EL MODO DE CALCULAR EL $R^2$ ..... 28
ECUACIÓN 4: FÓRMULA QUE REPRESENTA EL MODO DE CALCULAR EL CRPS ..... 29
ECUACIÓN 5: FÓRMULA QUE REPRESENTA EL MODO DE CALCULAR EL RMSD ..... 30
ECUACIÓN 6: FÓRMULA GENERAL DE REGRESIÓN LINEAL MÚLTIPLE ..... 64
ECUACIÓN 7:FÓRMULA GENERAL PARA CALCULAR LA POTENCIA PRODUCIDA PREVISTA, MODELO DE REGRESIÓN LINEAL MÚLTIPLE ..... 65
ECUACIÓN 8:FÓRMULA GENERAL PARA CALCULAR LA POTENCIA PREVISTA, MODELO DE REGRESIÓN LINEAL MÚLTIPLE ..... 67
ECUACIÓN 9:FÓRMULA PARAMETRIZADA CON LOS VALORES OBTENIDOS DE SOLVER, MODELO DE REGRESIÓN LINEAL MÚLTIPLE ..... 67

# 1. Resumen

En el presente trabajo final de grado, se realizará el estudio de diferentes modelos de predicción probabilísticos, los cuales serán aplicados a la potencia producida por una planta fotovoltaica situada en Alcolea del Río, provincia de Sevilla. Se tomarán como datos de partida, las predicciones meteorológicas de la ciudad de Alcolea del Río (datos obtenidos a través de MeteoGalicia) y la potencia producida por dicha planta fotovoltaica.

Los datos mencionados se utilizarán para predecir los valores de potencia que pudiera generar la planta fotovoltaica, para ello, la muestra deberá segmentarse en dos grupos: el primer grupo que corresponde con los datos de entrenamiento, que se utilizarán para elaborar y entrenar los diferentes modelos y el segundo grupo (que contendrá un número menor de datos) que corresponde con el grupo de test, el cuál permitirá evaluar los resultados obtenidos de la predicción y determinar su fiabilidad y precisión.

Para el desarrollo de los modelos de predicción probabilísticos se utilizarán dos programas informáticos, uno será la hoja de cálculo Excel y el otro será Rstudio.

Entre ambos permitirán generar el modelo climatológico, el modelo persistente probabilístico, los modelos de regresión y el modelo de bosques aleatorios. Una vez realizados todos los modelos se procederá a confeccionar una comparativa entre los modelos, determinando cuál ha generado mejores resultados en función de los criterios de evaluación explicados en dicho trabajo.

# 2. Abstract

In this final degree work, the study of different probabilistic prediction models will be carried out, which will be applied to the power produced by a photovoltaic plant located in Alcolea del Río, province of Seville. The meteorological forecasts of the city of Alcolea del Río (data obtained through MeteoGalicia) and the power produced by this photovoltaic plant will be taken as starting data.

The aforementioned data will be used to predict the power values that could be generated by the photovoltaic plant. To do this, the sample will have to be segmented into two groups: the first group corresponding to the training data, which will be used to elaborate and train the different models, and the second group (which will contain a smaller number of data) corresponding to the test group, which will make it possible to evaluate the results obtained from the prediction and determine their reliability and precision.

For the development of probabilistic prediction models, two computer programs will be used, one will be the Excel spreadsheet and the other will be Rstudio.

Between them, they will generate the climatological model, the probabilistic persistent model, the regression models and the random forest model. Once all the models have been carried out, a comparison will be made between the models, determining which has generated the best results based on the evaluation criteria explained in this work.

# 3. Introducción

Actualmente, la energía eléctrica es un bien de primera necesidad, se utiliza para una inmensidad de aplicaciones y es por ello que es imprescindible para el día a día, es decir, su utilidad va desde los hogares hasta comercios, industrias y multitud de ambientes.

Cabe destacar que, debido a diversos factores (cambio climático, aumento de la demanda eléctrica…) el sistema eléctrico ha sufrido cambios significativos a lo largo de los últimos años. Por tanto, el mercado eléctrico tal y como se conoce actualmente va a sufrir numerosos cambios; dichos cambios se producirán en gran parte por el cierre y desaparición de centrales térmicas y centrales nucleares y esto se debe a que las energías renovables están cobrando mayor importancia a causa del agravamiento del efecto invernadero y el consecuente calentamiento global.

Entre las energías renovables, uno de los tipos más destacados es la energía solar. Este tipo de energía renovable se obtiene a partir de la radiación electromagnética procedente del sol. Es importante recalcar que cada año, la radiación solar aporta a la Tierra la energía equivalente a varios miles de veces la cantidad de energía que se consume actualmente. Recogiendo la radiación solar de la manera adecuada, puede ser transformada en energía eléctrica o térmica mediante el uso de paneles fotovoltaicos.

La principal contrariedad que supone la utilización de energía solar, es que es un recurso intermitente y no determinista. Este problema puede provocar que la red eléctrica sea inestable, no obteniendo la producción necesaria para satisfacer la demanda de los consumidores (los posibles problemas que afectan a la producción de energía proveniente de la radiación solar, serán expuesto más adelante). Por tanto, es de primera necesidad poder predecir cuándo se van a producir estas situaciones que pueden disminuir la radiación solar.

A medida que se incrementa el número de plantas renovables integradas en la red general de distribución eléctrica, la gestión, mantenimiento y distribución de la electricidad y del sistema eléctrico en sí mismo, se vuelven más complejas por la intermitencia característica de este recurso, poniendo así en riesgo la estabilidad del sistema.

Para subsanar dicho problema, es muy importante el desarrollo de sistemas de predicción de radiación solar. Esto ayudaría a incrementar la cantidad de centrales de energía renovable sin complicar a su vez la gestión de la red eléctrica, a su vez; podría ser de gran interés la profundización en modelos de predicción de la energía eléctrica para las comercializadoras, de cara a las

compras en los mercados diario e intradiario. Es por ellos que se necesita un importante desarrollo en esta materia.

## 3.1 Objetivos del trabajo

El propósito final de este trabajo final de grado es la modelización y predicción probabilística a corto plazo de la energía eléctrica generada y vertida a la red por una planta fotovoltaica, teniendo a su vez en cuenta diferentes variables explicativas, como son la radiación media, temperatura, presión… junto con los datos históricos de la propia planta fotovoltaica con la finalidad de crear los distintos modelos de predicción probabilística.

Asimismo, cabe destacar que se ajustarán diferentes modelos, distinguiendo entre los que realizan una predicción probabilística de forma directa (utilizando los datos históricos de la propia planta fotovoltaica) y los que lo realizan de una forma indirecta (resultado de la mezcla del computo de variables). Los distintos modelos se desarrollarán mediante la hoja de cálculo Excel y el software libre R.

Una vez se hayan obtenidos los modelos de predicción probabilística y se hayan ajustado correctamente (evitando siempre que se pueda el sobre entrenamiento para que los errores que se produzcan en el testeo sean mínimos), se estimará la producción de energía eléctrica y se comparará entre los diferentes modelos, para poder así determinar cuál es el que proporciona una predicción probabilística más fiable, precisa y cuyos valores obtenidos se asemejen lo máximo posible a los reales.

Se establecerán los siguientes objetivos secundarios:

- Comprender y estudiar diferentes modelos matemáticos de predicción probabilística.
- Interpretar los datos obtenidos de cada uno de los modelos con la finalidad de determinar cuál se comporta mejor.
- Analizar los diferentes errores generados por los modelos de predicción.

### 3.2 Estructura del trabajo

En la realización de este proyecto se presentará de una manera breve la situación actual de la producción relacionada con las energías renovables a nivel internacional y en particular en España, centrándose a su vez en la energía solar.

Posteriormente, se procederá a la presentación de la planta fotovoltaica que se ha utilizado para llevar a cabo los diferentes modelos de predicción probabilística (planta fotovoltaica de Alcolea del Río, Sevilla), junto con los datos históricos recogidos de sus registros. Dichos datos, son fundamento indispensable para la realización y progreso del trabajo.

En los siguientes apartados se excluirán datos nulos o en blanco, así como errores procedentes de los diferentes aparatos de medición (se procede a la eliminación de "outliers" debido a que puede producir serios problemas en los modelos predictivos), se realizará dos grupos con la muestra (grupo de entrenamiento y grupo de test, cuyas finalidades se explicarán posteriormente).

Se procederá a tratar el objetivo principal de este trabajo, la explicación de los modelos de predicción confeccionados, una vez obtenidos los resultados pertinentes se procederá a realizar la comparación de los diferentes modelos (modelo climatológico, modelo de ventana fija, modelo de regresión lineal múltiple, modelo de regresión de cuantiles y modelo de bosques aleatorios) con la finalidad de obtener un pronóstico a corto plazo (entre veinticuatro y cuarenta y ocho horas) que será la base para el cálculo de energía eléctrica producida.

Concluyendo cuál es el modelo de predicción probabilística que proporciona menor error y se ajusta más a una producción real.

## 4. Situación actual de las energías renovables

Actualmente hay una gran dependencia energética. El problema real no es dicha dependencia, sino que a día de hoy la energía que utilizan casi todos los países, es energía procedente de los combustibles fósiles, los cuales generan una elevada contaminación y además contribuyen al aumento del efecto invernadero.

Asimismo, la presión social y la toma de conciencia de los gobiernos en la lucha contra el cambio climático, ha dado lugar a realizar cambios en los marcos reguladores con la pretensión de reducir los niveles actuales de emisión de $CO_2$.

Añadir a ello, que las energías renovables no han parado de crecer y después del carbón son las segundas fuentes globales de producción eléctrica.

Según la Agencia Internacional de Energía (AIE), la demanda mundial de electricidad aumentará un 70% en el año 2040, gracias a países de Oriente Medio y asiáticos.

Según el Informe del estado global de las renovables de 2018, las energías renovables ganan cada día más terreno a nivel mundial, pero el desarrollo es diferente según los sectores y regiones. En los países en desarrollo, por ejemplo, sobre todo en el África subsahariana, las tasas de acceso a la energía son bajas. En el año 2016 aproximadamente 1,06 billones de personas vivían sin electricidad. Los combustibles fósiles como el petróleo o el carbón, siguen siendo los más consumidos.

En cuanto a la situación en España, en base al Avance del Informe del sistema eléctrico español 2018 cabe destacar varios datos significativos:

- La producción de energía hidráulica crece un 85% respecto a 2017.
- La energía eólica se consolida como segunda fuente de generación de electricidad.
- La generación de energía renovable alcanza el 40,1%. De ese porcentaje el 49% corresponde a la energía eólica, el 34% a la hidráulica, el 11% a la solar y el 5% a otros.

A través de los datos expuestos, se puede deducir que en España la importancia de la integración y el desarrollo de las energías renovables para la producción de electricidad está creciendo considerablemente.

## 4.1 Situación de la energía solar fotovoltaica en España

Según los datos registrados por UNEF (Unión Española Fotovoltaica), la reactivación del sector solar fotovoltaico nacional es vertiginosa: 55 MW en 2016; 135 MW en 2017; y 261.7 MW en 2018. De estos, un 90% aproximadamente (235.7 megavatios) corresponde al autoconsumo energético y 26 MW a plantas solares fotovoltaicas sobre suelo (en el segmento del autoconsumo, el 25% del total fueron instalaciones para uso agrícola conectadas a la red).

En el siguiente gráfico, se muestra la potencia solar fotovoltaica instalada en España desde el 2006.

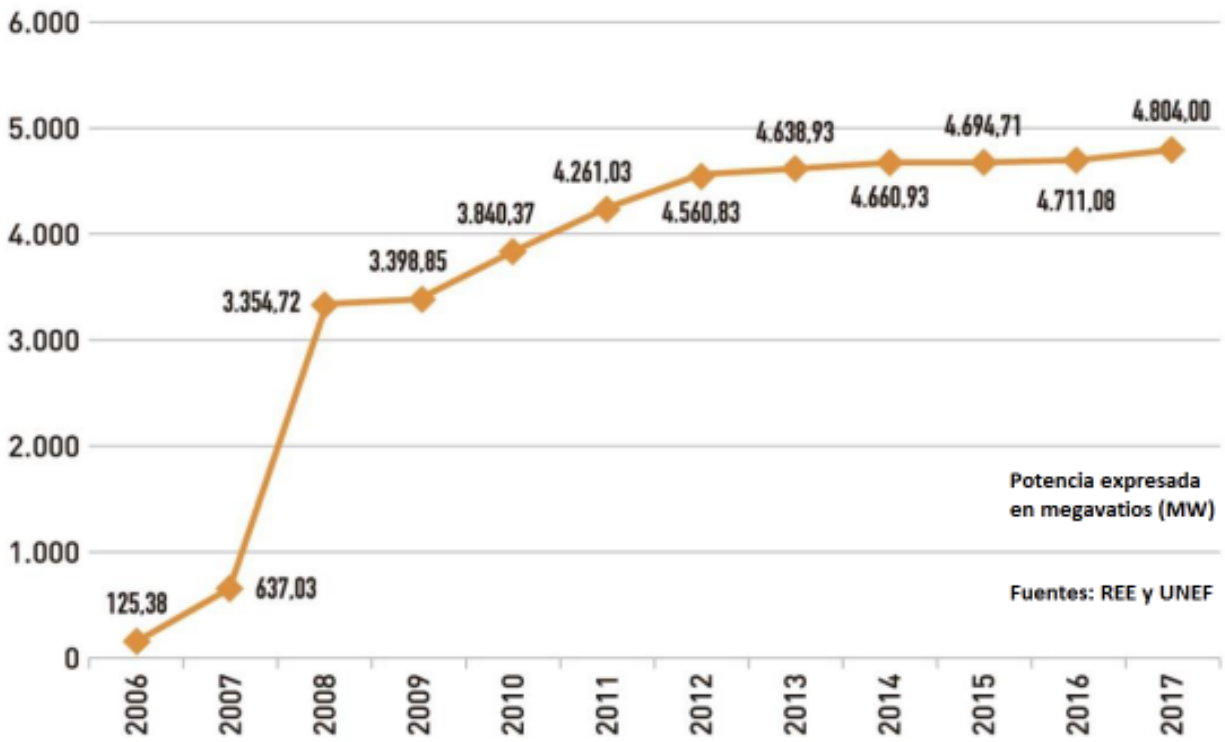


*Ilustración 1: Potencia solar fotovoltaica instalada acumulada en España*

Cabe destacar que, aunque se haya producido cierto crecimiento, la cifra de potencia instalada en España únicamente representa el 3% de la nueva potencia instalada en Europa en 2018, con Alemania y Países bajos como los principales precursores de la instalación de energía fotovoltaica.

La UNEF determina que algunas de las razones por las que se ha producido dicho crecimiento son las siguientes:

- Reducción en los costes de producción en los últimos años.
- Impulso de la directiva europea de Energías renovables.
- Derogación del impuesto al sol.
- Ahorro energético y optimización financiera (orientado en mayor parte a las empresas).

En la gráfica que se presentará a continuación, se puede ver la tendencia a nivel mundial del crecimiento de la potencia solar fotovoltaica instalada en todo el mundo entre los años 2000 y 2017.

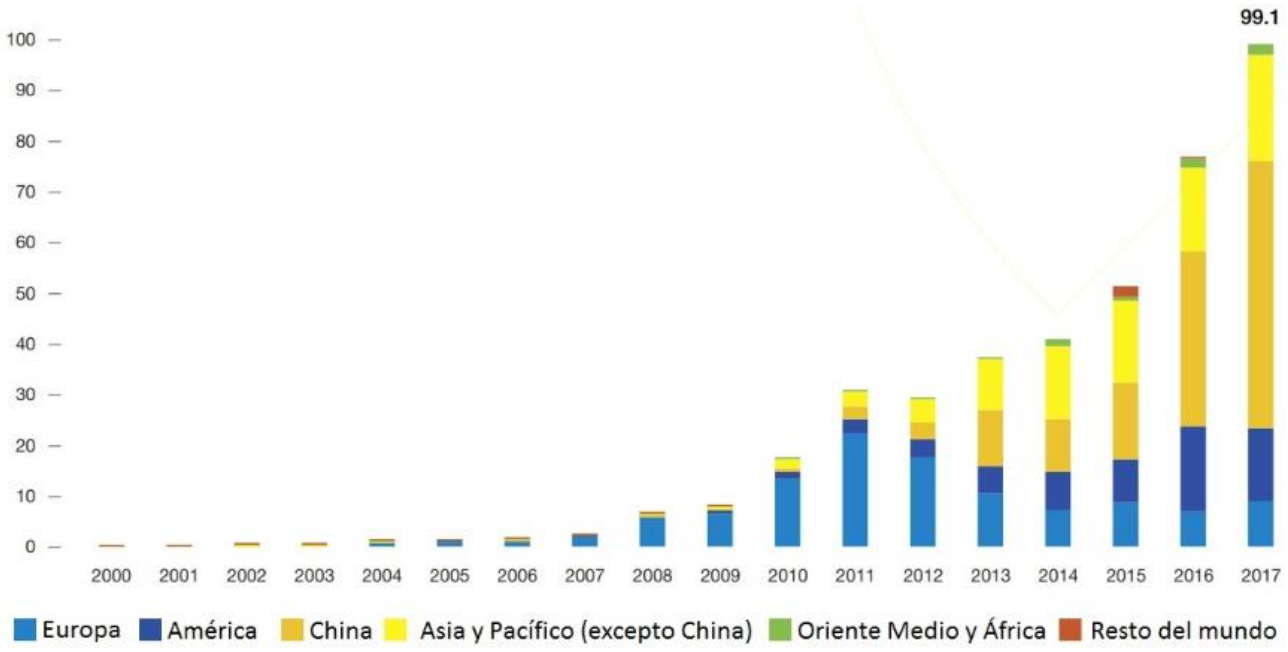


*Ilustración 2: Potencia solar fotovoltaica instalada anualmente en todo el mundo entre los años 2000 y 2017 (expresada en GW)*

Como conclusión, se puede decir que en España aún queda mucho por hacer, y si países situados en esas latitudes han conseguido hacer que la energía solar fotovoltaica sea rentable y hayan apostado por su desarrollo; en España se podría conseguir una mayor rentabilidad y producción debido a que se disponen de más horas de sol a lo largo del año. A su vez, se observa que hay una clara apuesta, a nivel mundial, por el desarrollo fotovoltaico, pretendiendo que las energías renovables se conviertan en la única fuente de energía eléctrica para un horizonte temporal de unos 30 años.

# 5. Necesidad de la predicción de la energía solar

En primer lugar, hay que remarcar que el problema más importante vinculado con la producción de energía eléctrica, es la imposibilidad de almacenar la energía a gran escala. Asimismo, otra de las grandes funciones que debe garantizar un sistema eléctrico, es poder cubrir la demanda de energía que soliciten los diferentes usuarios, estando disponible en todo momento y con suministro constante (a excepción de posibles bajadas o subidas de tensión producidas en lapsos de tiempo muy cortos y de forma esporádica).

Por ello, la primera variable a tener en cuenta dentro de este problema, es la demanda energética, es decir, el poder predecir la demanda de energía eléctrica, permitirá realizar una planificación adecuada en las centrales, en comercializadoras de energía eléctrica (podrá influir en la determinación de las tarifas eléctricas) y en el Sistema Eléctrico de forma general, del país.

Teniendo en cuenta lo comentado anteriormente, la energía eléctrica producida mediante plantas fotovoltaicas siempre se ha considerado una fuente de producción de baja fiabilidad, cuyo régimen de generación de energía se produce con alta intermitencia (esto es debido a que dicha producción depende en gran parte de las condiciones climatológicas, sobre todo de la cobertura de nubes) y la contrariedad de no poder tener control de la producción eléctrica (depende de la radiación aportada por el sol).

Por tanto, aparece la necesidad intrínseca de predecir la energía eléctrica que se produce en las plantas fotovoltaicas. Asimismo, el principal objetivo de las predicciones a corto plazo es conocer anticipadamente y lo más exacto posible cuánta energía va a generar el sistema para que puedan participar en los despachos de energía realizados por los operadores del sistema eléctrico.

Este criterio de asignación se basa tanto en criterios técnicos que buscan asegurar la fiabilidad y seguridad del suministro eléctrico como en criterios económicos que persiguen reducir los costes de generación de energía eléctrica. Los despachos de electricidad se realizan a escala horaria, para un día completo y con un mínimo de 12 horas de adelanto.

A continuación, se va explicar brevemente por qué actualmente los modelos de predicción probabilísticos están ganando peso frente a los modelos de predicción deterministas.

Los primeros modelos en desarrollarse son de tipo determinístico (también llamados puntual o spot). Son aquellos donde se supone que los datos se conocen con certeza, es decir, donde las mismas entradas originarán invariablemente las mismas salidas, no contemplándose la existencia del azar ni el principio de incertidumbre. Por tanto, solo dan el valor esperado de la generación, sin ninguna otra información acerca de lo buena o mala que puede ser esa predicción.

Asimismo, en los últimos años han aparecido los modelos probabilísticos, que proporcionan la forma que pueden tomar un conjunto de datos obtenidos del muestreo de valores con un comportamiento que se supone aleatorio, además del valor esperado y la posible distribución estadística del error de predicción (diferencia entre el valor esperado y el real). Este tipo de modelo de predicción proporciona una información completa que puede ser de gran utilidad para

evaluar el riesgo económico en operaciones en el mercado (venta de la energía generada), permitiendo analizar de antemano las probabilidades de obtener o no los beneficios perseguidos (por el tema de las penalizaciones en el mercado eléctrico por la no generación de los valores de energía ofertados).

# 6. Horizontes de predicción

Para elaborar un modelo de predicción de producción de energía eléctrica, se debe contemplar como uno de los factores más importantes el horizonte de predicción, es decir, el periodo de tiempo que determina el momento del futuro y que está íntimamente ligado a la variable o variables a predecir. Por tanto, se diferencian tres tipos de horizontes de predicción:

- **Predicciones a corto plazo:** este horizonte generalmente se separa en dos subgrupos: muy corto plazo y corto plazo.

  Las predicciones a muy corto plazo, se emplea fundamentalmente como medio para la regulación de los flujos de carga, a su vez estas predicciones son de gran interés para el operador del sistema, permitiéndole llevar a cabo un adecuado funcionamiento y mantenimiento del sistema eléctrico. El horizonte de predicción en este caso es de unas horas después de la predicción del consumo.

  El horizonte a corto plazo, abarca desde el límite del horizonte de muy corto plazo hasta unas 48 ó 72 horas. Se podría decir que este horizonte de predicción es el más interesante, debido a que influye de forma directa en el mercado diario eléctrico.

- **Predicciones a medio plazo:** este horizonte de predicción suele abarcar entre un mes y el año. Para su realización, los parámetros de partida son las mediciones existentes de consumidores con una frecuencia de dos meses.

- **Predicciones a largo plazo:** es el horizonte de previsión de demanda de mayor duración, superando el año. Cabe destacar, que cuanto mayor sea el horizonte de predicción, mayor será el error que se generará en la predicción. Las predicciones realizadas a largo plazo son de bastante utilidad para controlar el equilibrio entre la demanda y la generación relacionándolo con la situación actual del país.

A falta de horizontes tan bien delimitados en el campo de la predicción en plantas fotovoltaicas como en el campo de la demanda de energía eléctrica, utilizaré esa misma terminología y plazos en este Trabajo Fin de Grado. Es decir, denominaré como horizonte a corto plazo como el correspondiente a las siguientes 72 horas, aunque con especial interés por el que comprende las 24 horas del día siguiente al del momento de realizar la predicción.

# 7. Herramientas utilizadas para el desarrollo de los modelos de predicción

## 7.1 Programa Microsoft Excel

Excel es una herramienta ofimática perteneciente al conjunto de programas denominados hoja de cálculo electrónica, en la cual se puede escribir, almacenar, manipular, calcular y organizar todo tipo de información numérico o de texto.

Excel es una hoja de cálculo electrónica desarrollado por Microsoft, el cual se encuentra dentro del paquete de herramientas o programas ofimáticos llamados Office, el programa ofimático Excel es la hoja de cálculo electrónica más extendida y usada a nivel global, hoy en día el trabajo de cualquier ingeniero, financiero, matemático, físico o contable sería muy diferente sin la aplicación de cálculo Excel.

La principal ventaja del programa Excel es la versatilidad y funcionalidad que presenta a la hora de realizar cualquier tipo de modelo, con Excel podemos generar hojas Excel para el diseño y cálculo de estructuras civiles, gestión y control de la contabilidad de una empresa, gestión y control de los stocks de un almacén, diseños de modelos matemáticos, gestión de bases de datos, generación de presupuestos, planificación de proyectos, etc.… Un amplio abanico de posibilidades se puede cubrir con el uso del programa Excel.

## 7.2 Programa R

R es un entorno y lenguaje de programación orientado a objetos para análisis estadístico y gráfico. R es software libre, resultado de la implementación GNU del lenguaje S. R es el lenguaje más utilizado en investigación estadística, y además cuenta con una popularidad especial en los campos de la investigación biomédica, las matemáticas financieras y la bioinformática.

El software libre R, ofrece una amplia gama de herramientas estadísticas (como modelos lineales, modelos no lineales, análisis de series temporales, algoritmos de clasificación, etc.) y herramientas gráficas para representación de datos.

Las principales ventajas que ofrece R son las siguientes:

- R se distribuye bajo licencia GNU. Esto quiere decir que su utilización es completamente libre y gratuita.
- R es multiplataforma. Eso implica que no sólo está disponible para los sistemas operativos Windows, GNU/Linux, Unix y Macintosh, también puede integrarse con distintas bases de datos y se puede utilizar desde lenguajes de programación interpretados como Python.
- Se puede utilizar cualquier tipo de datos. Tiene la gran ventaja de ser compatible con todos los formatos de datos conocidos (CSV, XLS, etc.).
- Su capacidad gráfica permite generar gráficos de alta calidad. Cualquier otro paquete estadístico no supera dicha capacidad.

### 7.2.1 La interfaz Rstudio

RStudio es un Entorno de Desarrollo Integrado (IDE) para R. RStudio está compuesto por cuatro áreas de trabajo diferentes. En la siguiente ilustración, puede verse la interfaz donde se aprecian dichas áreas:

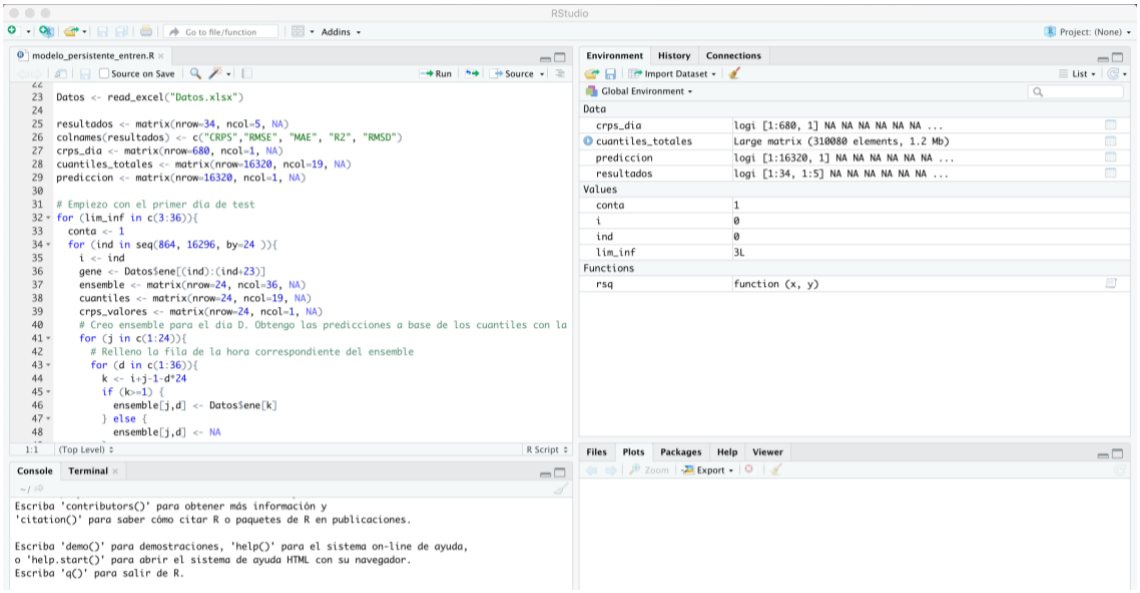


*Ilustración 3: muestra entorno de trabajo de Rstudio*

En el área superior izquierda se pueden abrir, crear y editar ficheros, generalmente scripts con código R, aunque también se puede trabajar con otro tipo de ficheros.

En el área inferior izquierda, se encuentra ubicad la consola de comandos de R. Desde el área superior izquierda también puede seleccionarse partes de scripts o un script entero y enviarlo a ejecutar en dicha consola.

En el área superior derecha se sitúan dos pestañas:

- **Workspace:** Contiene la lista de objetos creados en memoria.
- **History:** Contiene el historial de líneas de código ejecutadas en la consola de R.

Para concluir, el área inferior derecha está compuesta por diversas pestañas:

- **Files:** Da acceso al árbol de directorios y ficheros del disco duro.
- **Plots:** En esta pestaña, aparecen los gráficos que se crean en la consola mediante comandos. Aparecen varias opciones, como hacer zoom en los gráficos o exportarlos a un archivo.
- **Packages:** Desde esta pestaña se realiza la administración de los paquetes de R que están instalados en la máquina. A través de aquí se puede acceder al repositorio de paquetes de R para instalar los paquetes que sean necesarios y no estén instalados.
- **Help:** En esta pestaña se puede acceder a toda la documentación oficial de R. Si ejecutamos en la consola la ayuda de un comando de R, la página de ayuda de dicho comando se abre aquí.
- **Viewer:** Panel que puede ser utilizado para ver contenido web local.

# 8. Desarrollo del trabajo

## 8.1 Planta fotovoltaica de Alcolea del Río

La planta fotovoltaica elegida para desarrollar los diferentes modelos de predicción, se encuentra situada en Alcolea del Río, provincia de Sevilla y es una planta fotovoltaica de 2160 kW pico. El proyecto se encuentra en una de las zonas de mayor insolación de la península. La finca en la cual se localiza la instalación, tiene una extensión de 8.6 hectáreas aproximadamente y es completamente llana. El diseño previsto está planteado para un impacto mínimo sobre el terreno.

Los componentes que constituyen dicha instalación son los siguientes:

- Estructura: superficie sobre la que se instala el generador fotovoltaico para que quede con la orientación e inclinación más viable. En este caso, se han usado estructuras a 25º de inclinación y orientación Sur, de acero galvanizado en caliente con pilares hincados directamente en el terreno.

- Módulos fotovoltaicos: son los encargados de captar la radiación solar y producir electricidad en corriente continua. Se ha instalado una potencia total de 2160 kWp distribuida en 8.640 módulos Exiom de 250 Wp cada uno.

- Inversores: es el elemento que transforma la energía eléctrica en corriente continua en energía eléctrica en corriente alterna, para poderla verter a la red eléctrica general y así venderla al proveedor energético. Para la conexión a la red se han instalado los siguientes elementos:

    1. Casetas de inversores: se han instalado 2 casetas de inversores de 1MW cada una con centro de transformación incluido. Cada caseta dispone de 2 inversores de 500kW.

    2. Centro de Seccionamiento: la energía producida de las casetas de inversores (en media tensión) se lleva hasta el centro de seccionamiento y de este se inyecta en la red. Dicho centro de seccionamiento se cede a la compañía eléctrica.

A continuación, se podrá apreciar la distribución final de la planta fotovoltaica:

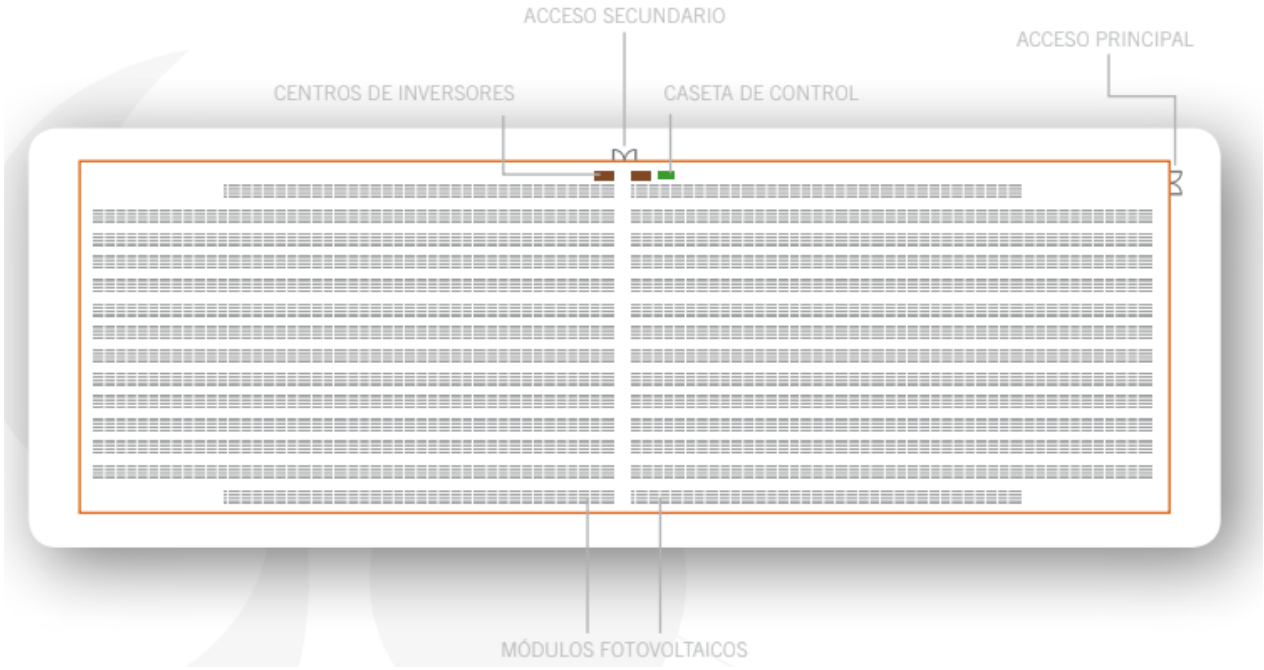


*Ilustración 4: Representación desde vista aérea de la distribución de los módulos fotovoltaicos*

*Ilustración 5: Representación real de los módulos fotovoltaicos de Alcolea del Río*

## 8.2 Realización de la muestra

Para el desarrollo de los modelos de predicción de la producción de energía solar fotovoltaica se han utilizado dos bases de datos, las cuales son las siguientes:

- Potencias producidas por la planta fotovoltaica de Alcolea del Río, provincia de Sevilla, potencias producidas entre el 1 de junio de 2016 hasta el 9 de marzo de 2019.
- Predicciones meteorológicas para la ciudad de Alcolea del Río, situada en la provincia de Sevilla, con predicción de hasta los cuatro días siguientes al comienzo de la toma de datos el 1 de junio de 2016 hasta el 9 de marzo de 2019.

### 8.2.1 Predicciones meteorológicas

Se puede deducir que la potencia producida en una planta fotovoltaica está íntimamente relacionada e influida por la meteorología. Lo que se pretende explicar con ello, es que la relación existente entre la potencia producida por una planta fotovoltaica está influenciada en gran medida por factores climatológicos externos como son la temperatura exterior, la humedad relativa, la radiación, la cobertura de nubes, etc. Asimismo, se han seleccionado algunas variables meteorológicas como variables predictivas de la producción de energía.

Por tanto, se dispone de las predicciones meteorológicas para la ciudad de Alcolea del Río, comenzando dichas predicciones el 1 de junio de 2016 hasta el 9 de marzo de 2019, se disponen de las siguientes previsiones: 1 día, 2 días, 3 días y 4 días, utilizando para los modelos de predicción probabilísticos la predicción para un día vista, es decir, del día D+1. Las predicciones han sido obtenidas del Servidor Meteorológico de Galicia (MeteoGalicia), organismo que genera las predicciones mediante la utilización de programas de predicción numérica meteorológica (NWP), cuyo funcionamiento se basa principalmente en modelar el espacio que se desea analizar y lo convierte en una rejilla de análisis, por ejemplo, cogiendo puntos distanciados entre sí 20 km (en superficie) y diferentes catas de altura, por ejemplo, a 2 metros, 5 metros, 100 metros… obteniendo el comportamiento de la atmósfera.

En MeteoGalicia utilizan el modelo WRF (**Weather Research and Forecasting**), que es un sistema de cálculo numérico para simulación atmosférica diseñado para satisfacer las necesidades tanto de investigación como de predicción atmosféricas. WRF incluye dos núcleos diferentes (ARW, NMM), un sistema de asimilación de datos, y una arquitectura de software diseñada para la posibilidad de ejecuciones distribuidas o paralelas y la escalabilidad del sistema. WRF implementa una extensa gama de aplicaciones meteorológicas en escalas que van desde los metros a los miles de kilómetros.

El proceso se basa en la utilización de dominios anidados, se utilizarán tres dominios (d1, d2 y d3 los cuales se mostrarán de forma gráfica en la siguiente página) sobre los cuales se realizan las predicciones meteorológicas.

Primero se hacen las predicciones sobre el dominio 1 (la separación entre los puntos en este caso serán unos 25 km), después sobre el dominio 2 (la separación entre los puntos en este caso serán unos 12 km) y por último en el dominio 3 (la separación entre los puntos en este caso serán unos 3 km). Para la planta fotovoltaica de Alcolea del Río, se ha escogido el dominio 2 porque proporciona una mayor precisión de los datos que se obtendrán.


*Ilustración 6: Dominios anidados que delimitan sobre que área se realizarán las predicciones meteorológicas*

Conociendo la longitud y latitud de Alcolea del Río (longitud: -5,64431 y latitud: 37,634030), se determinaron cuales eran los cuatro puntos más cercanos para poder realizar mediante interpolación cuadrática la predicción meteorológica de las variables de interés para los modelos de predicción probabilística.


*Ilustración 7: Situación de los 4 puntos más cercanos a Alcolea del Río para hacer la interpolación cuadrática de la predicción meteorológica*

A continuación, se mostrarán las diferentes predicciones meteorológicas que se han utilizado para la realización de los modelos de predicción:

- Precipitación
- Fracción total de nubes
- Humedad relativa
- Presión
- Radiación media
- Temperatura
- Velocidad de viento

### 8.2.2 Potencias producidas por la planta fotovoltaica

Para la creación de los modelos de predicción de la potencia producida se ha empleado el histórico de datos de la planta fotovoltaica de Alcolea del Río. Dichos datos, corresponden al periodo de tiempo entre el 1 de junio de 2016 hasta el 9 de marzo de 2019, lo que supone un total de 1013 días. Los valores recopilados para el desarrollo de los modelos de predicción fueron registrados con una periodicidad de 60 minutos.

La muestra ha sido dividida en dos grupos:

- **Grupo 1:** corresponde con los datos de entrenamiento, que se utilizarán para realizar el ajuste o entrenamiento del modelo, optimizando diferentes parámetros intentando reducir lo máximo posible el error de la predicción, lo que se traduce en que a menor

error más preciso será el modelo realizado y, por tanto, mayor exactitud existirá entre las entradas y la variable de salida.

- **Grupo 2:** corresponde con los datos de test, utilizados para determinar la precisión de los modelos creados, es decir, se utilizarán diferentes criterios (los cuales se especificarán más adelante) para determinar la adecuación de las predicciones a la realidad.

En primer lugar, para realizar un uso adecuado de la muestra, se deberá realizar un primer "barrido" de los datos con la intención de determinar la ausencia de datos o la aparición de valores atípicos que influyan de forma negativa en la realización de los diferentes modelos.

Asimismo, la depuración de la muestra permite detectar o en su defecto corregir los posibles valores erróneos de una base de datos. Dicho proceso se utiliza especialmente cuando en alguna parte existen datos incorrectos, incompletos que posteriormente han sido modificados, sustituidos o eliminados.

A continuación, tras haber realizado una primera depuración se realizará una representación de las potencias producidas (se ha realizado la representación de los dos grupos de la muestra para determinar la existencia de algún defecto en alguno de ellos) para evaluar si existe algún tipo de problema en la recolección de los datos, como falta de datos, "outliers" u otros posibles defectos.

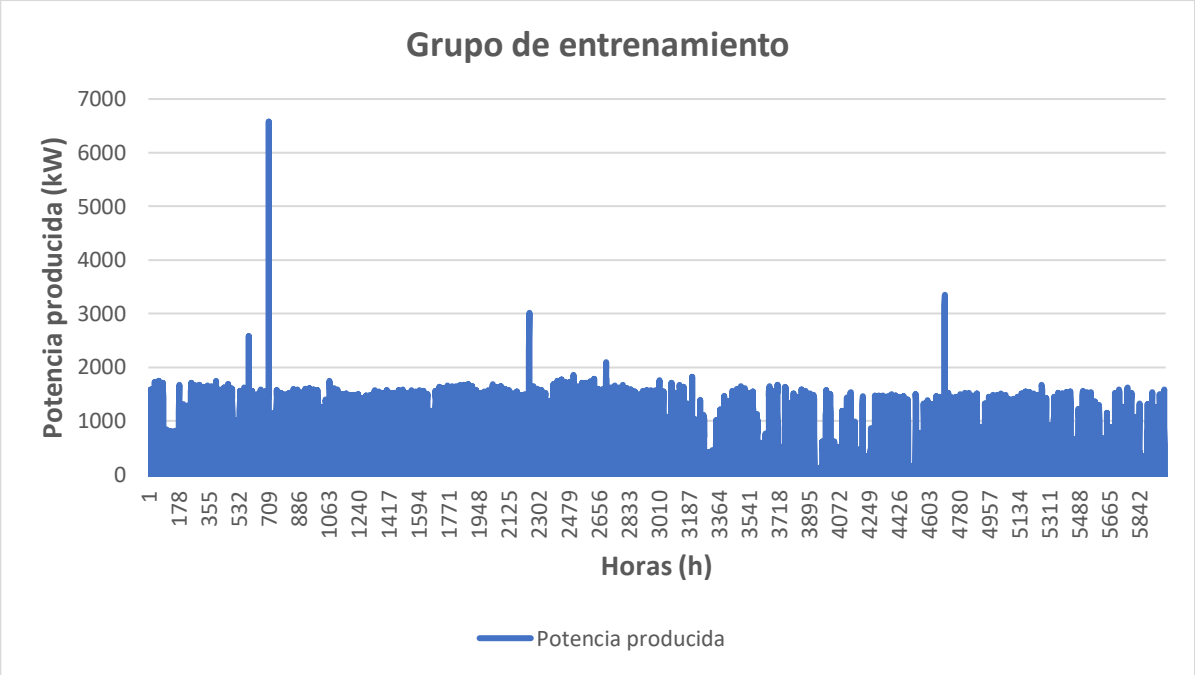


*Ilustración 8: Representación de la muestra correspondiente al grupo de entrenamiento con "outliers"*

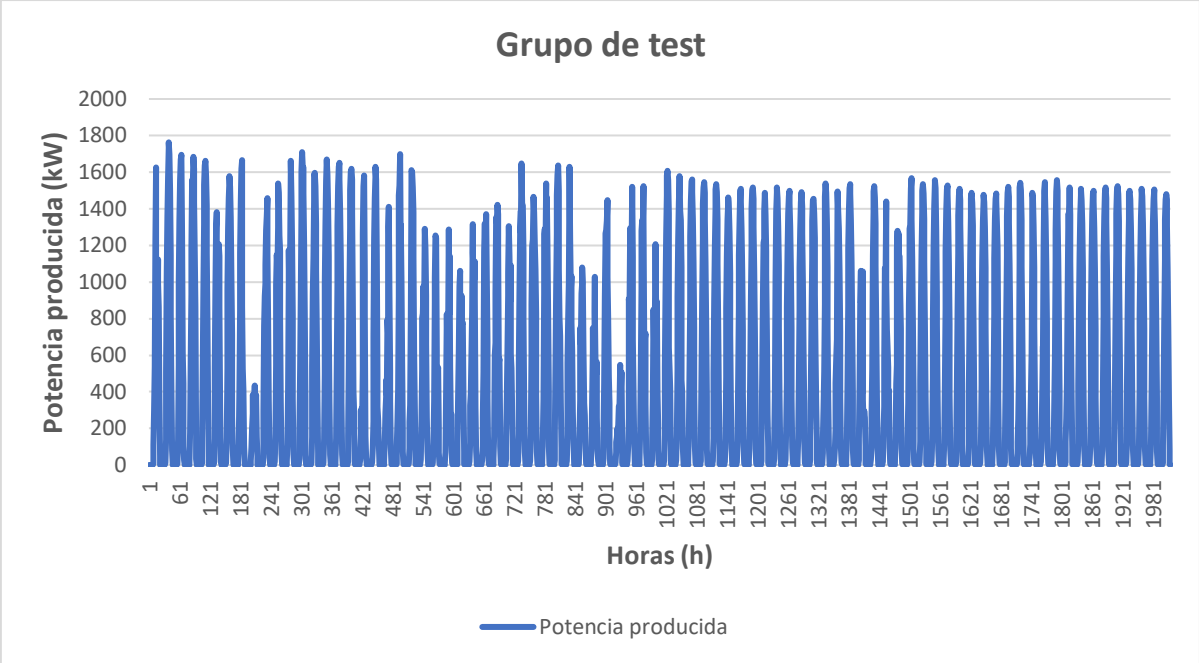


*Ilustración 9: Representación de la muestra correspondiente al grupo de test*

De las figuras anteriores, se deduce que, en la muestra correspondiente al grupo de entrenamiento, existen todavía ciertos "outliers" que deberán ser eliminados con la finalidad de que no afecten a los resultados que se obtengan en los modelos de predicción.

Respecto al grupo de test, no se detecta ninguna anomalía o valor atípico reseñable, y no habrá que hacer ninguna depuración más sobre este grupo.

Una vez se han eliminado los "outliers" existentes en el grupo de entrenamiento, quedan las dos muestras completas, aunque los valores correspondientes al grupo de entrenamiento se verán reducidos. Por tanto, el grupo 1 estará compuesto finalmente con 16318 valores y el grupo 2 estará constituido por 3673 valores.

En la siguiente tabla, se representa un ejemplo de los datos con una periodicidad de una hora en la que se muestran las potencias producidas, la hora civil, hora GMT y las diferentes predicciones meteorológicas de las que se disponen para utilizar en los modelos de predicción.

| Fecha | Hora Civil | Potencia producida (kW) | Hora GMT | Temperatura (k) | Radiación (W/m2) | Rad_m (W/m2) | Presión (Pa) | Velocidad (m/s) | Humedad rel | Fraccion total nubes | Precipitación (mm) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2016-06-02 | 02:00 | 0 | 00:00:00 | 290,108661 | 0 | 0 | 101588,2693 | 2,307659971 | 0,7726754 | 0 | 0 |
| 2016-06-02 | 03:00 | 0 | 01:00:00 | 289,3268661 | 0 | 0 | 101575,8199 | 2,750492541 | 0,7701647 | 0 | 0 |
| 2016-06-02 | 04:00 | 0 | 02:00:00 | 288,4524355 | 0 | 0 | 101570,3402 | 2,634815699 | 0,7949411 | 0,003219396 | 0 |
| 2016-06-02 | 05:00 | 0 | 03:00:00 | 287,7662542 | 0 | 0 | 101535,4959 | 2,639697476 | 0,7994337 | 0,086417092 | 0 |
| 2016-06-02 | 06:00 | 0 | 04:00:00 | 287,237258 | 0 | 0 | 101485,0717 | 2,789303223 | 0,7950789 | 0,15367125 | 0 |
| 2016-06-02 | 07:00 | 0 | 05:00:00 | 286,8809442 | 0 | 0 | 101525,0038 | 3,086823737 | 0,7853915 | 0,163116385 | 0 |
| 2016-06-02 | 08:00 | 3 | 06:00:00 | 287,005727 | 34,87833877 | 17,43916938 | 101564,2995 | 3,28854671 | 0,778315 | 0,134114001 | 0 |
| 2016-06-02 | 09:00 | 59 | 07:00:00 | 289,6042394 | 275,5851742 | 155,2317565 | 101562,1149 | 3,561701579 | 0,6782033 | 0,087284131 | 0 |
| 2016-06-02 | 10:00 | 603 | 08:00:00 | 292,203501 | 547,2739735 | 411,4295738 | 101551,5414 | 4,008678288 | 0,6171052 | 0,05797769 | 0 |
| 2016-06-02 | 11:00 | 1025 | 09:00:00 | 294,5441688 | 663,2163998 | 605,2451866 | 101534,7683 | 3,81167646 | 0,5785177 | 0,032738887 | 0 |
| 2016-06-02 | 12:00 | 1356 | 10:00:00 | 296,84229 | 834,8629378 | 749,0396688 | 101518,2484 | 3,422682991 | 0,5264993 | 0,176949774 | 0 |
| 2016-06-02 | 13:00 | 1556 | 11:00:00 | 298,6911628 | 869,9320735 | 852,3975057 | 101504,2331 | 1,886487553 | 0,4630992 | 0,201028001 | 0 |
| 2016-06-02 | 14:00 | 1488 | 12:00:00 | 300,5420674 | 957,7545883 | 913,8433309 | 101450,1503 | 1,906133608 | 0,412402 | 0,224989169 | 0 |
| 2016-06-02 | 15:00 | 1582 | 13:00:00 | 301,9056989 | 944,8582807 | 951,3064345 | 101386,7705 | 1,328280463 | 0,3556084 | 0,156418668 | 0 |
| 2016-06-02 | 16:00 | 1473 | 14:00:00 | 302,577866 | 819,2772829 | 882,0677818 | 101306,3862 | 1,133309328 | 0,3257555 | 0,270985779 | 0 |
| 2016-06-02 | 17:00 | 1301 | 15:00:00 | 302,9467706 | 836,5112613 | 827,8942721 | 101273,1264 | 1,436241583 | 0,3184476 | 0,014652474 | 0 |
| 2016-06-02 | 18:00 | 587 | 16:00:00 | 302,3335353 | 366,6401212 | 601,5756913 | 101233,0134 | 1,781798286 | 0,3529002 | 0,461701656 | -1,36442E-08 |
| 2016-06-02 | 19:00 | 327 | 17:00:00 | 301,3668515 | 199,1219587 | 282,8810399 | 101188,101 | 2,086751986 | 0,4270186 | 0,200000003 | -2,8687E-09 |
| 2016-06-02 | 20:00 | 116 | 18:00:00 | 300,0544795 | 174,8503342 | 186,9861465 | 101161,7435 | 2,244565326 | 0,4997047 | 0,430011766 | -2,65821E-09 |
| 2016-06-02 | 21:00 | 44 | 19:00:00 | 297,3267546 | 8,821439665 | 91,83588694 | 101135,1864 | 1,986519501 | 0,6416585 | 0,141118131 | 0 |
| 2016-06-02 | 22:00 | 13 | 20:00:00 | 295,6873257 | 0 | 4,410719833 | 101177,7448 | 2,271942938 | 0,7205873 | 0,178422891 | 0 |
| 2016-06-02 | 23:00 | 0 | 21:00:00 | 293,7739458 | 0 | 0 | 101230,7719 | 2,063770154 | 0,8162842 | 0,163643266 | 0 |
| 2016-06-03 | 00:00 | 0 | 22:00:00 | 292,8770189 | 0 | 0 | 101273,2771 | 1,6316833 | 0,8421519 | 0,200000003 | 0 |
| 2016-06-03 | 01:00 | 0 | 23:00:00 | 292,0154206 | 0 | 0 | 101248,447 | 1,700870487 | 0,8730832 | 0,200000003 | 0 |
| . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . |
| 2018-09-30 | 00:00 | 0 | 22:00:00 | 295,6925928 | 0 | 0 | 101655,7621 | 1,528988898 | 0,8546782 | 0 | 0 |
| 2018-09-30 | 01:00 | 0 | 23:00:00 | 294,9916683 | 0 | 0 | 101631,965 | 1,581082307 | 0,8704006 | 0 | 0 |
| 2018-09-30 | 02:00 | 0 | 00:00:00 | 294,0728798 | 0 | 0 | 101520,3929 | 0,565761245 | 0,9057936 | 0 | 0 |
| 2018-09-30 | 03:00 | 0 | 01:00:00 | 293,5627733 | 0 | 0 | 101519,0685 | 0,703522601 | 0,9314007 | 0 | 0 |
| 2018-09-30 | 04:00 | 0 | 02:00:00 | 293,1312235 | 0 | 0 | 101520,9471 | 0,866608822 | 0,9507126 | 0 | 0 |
| 2018-09-30 | 05:00 | 0 | 03:00:00 | 292,768975 | 0 | 0 | 101514,548 | 1,296717889 | 0,9659855 | 0 | 0 |
| 2018-09-30 | 06:00 | 0 | 04:00:00 | 292,4733048 | 0 | 0 | 101529,1132 | 1,148011776 | 0,9752943 | 0,055030348 | 0 |
| 2018-09-30 | 07:00 | 0 | 05:00:00 | 292,2309582 | 0 | 0 | 101558,3916 | 0,863345329 | 0,9793939 | 0,142580128 | 0 |
| 2018-09-30 | 08:00 | 0 | 06:00:00 | 292,0464309 | 0 | 0 | 101583,0415 | 0,670351499 | 0,9819152 | 0,094047309 | 0 |
| 2018-09-30 | 09:00 | 27 | 07:00:00 | 292,3085862 | 43,9740041 | 21,98700205 | 101623,1232 | 0,936979789 | 0,9653818 | 0,037084712 | 0 |
| 2018-09-30 | 10:00 | 267 | 08:00:00 | 293,5513737 | 160,498387 | 102,2361956 | 101670,0471 | 1,69346604 | 0,8969867 | 0,016454626 | 0 |
| 2018-09-30 | 11:00 | 690 | 09:00:00 | 295,50006 | 403,6323049 | 282,065346 | 101705,4093 | 1,620203692 | 0,813269 | 0 | 0 |
| 2018-09-30 | 12:00 | 1048 | 10:00:00 | 297,4442411 | 601,2300331 | 502,431169 | 101718,2654 | 1,607817653 | 0,728488 | 0 | 0 |
| 2018-09-30 | 13:00 | 1339 | 11:00:00 | 299,2369863 | 673,7978667 | 637,5139499 | 101685,9836 | 1,523841315 | 0,6499714 | 0 | 0 |
| 2018-09-30 | 14:00 | 1396 | 12:00:00 | 300,8327112 | 753,7405771 | 713,7692219 | 101668,9645 | 1,148794755 | 0,5720649 | 0 | 0 |
| 2018-09-30 | 15:00 | 1460 | 13:00:00 | 301,8206421 | 740,2791128 | 747,0098449 | 101625,9499 | 0,89364728 | 0,5175874 | 0 | 0 |
| 2018-09-30 | 16:00 | 1231 | 14:00:00 | 302,4403276 | 698,6995854 | 719,4893491 | 101554,872 | 0,970437141 | 0,4843504 | 0 | -1,01014E-08 |
| 2018-09-30 | 17:00 | 1255 | 15:00:00 | 302,6498437 | 551,6826941 | 625,1911398 | 101493,918 | 1,115508758 | 0,467569 | 0 | 0 |
| 2018-09-30 | 18:00 | 876 | 16:00:00 | 302,4839869 | 336,1120732 | 443,8973837 | 101445,1566 | 1,31529578 | 0,4736056 | 0 | 0 |
| 2018-09-30 | 19:00 | 418 | 17:00:00 | 302,0309648 | 212,8938167 | 274,502945 | 101417,7004 | 1,612468532 | 0,5149212 | 0 | 0 |
| 2018-09-30 | 20:00 | 254 | 18:00:00 | 299,9737327 | 1,791945412 | 107,342881 | 101437,5152 | 1,633545294 | 0,6407963 | 0 | 0 |
| 2018-09-30 | 21:00 | 0 | 19:00:00 | 298,4272907 | 0 | 0,895972706 | 101505,7992 | 1,886254204 | 0,726508 | 0 | 0 |
| 2018-09-30 | 22:00 | 0 | 20:00:00 | 297,4097943 | 0 | 0 | 101577,2063 | 2,333193471 | 0,7446327 | 0 | 0 |
| 2018-09-30 | 23:00 | 0 | 21:00:00 | 296,5918133 | 0 | 0 | 101615,9061 | 1,91519427 | 0,7949961 | 0 | 0 |

*Tabla 1: Ejemplo de las muestras utilizadas para el desarrollo de los modelos de predicción*

# 9. Criterios de evaluación de los modelos de predicción

## 9.1 Modelos determinísticos o puntuales

Los modelos determinísticos o puntuales permiten evaluar principalmente la precisión del modelo elaborado, es decir, la precisión cuantifica la desviación de los valores pronosticados frente a los valores reales.

Por tanto, los indicadores estadísticos utilizados normalmente en modelos de predicción determinísticos que permiten estimar el adecuado funcionamiento de un modelo de predicción son, el error absoluto medio, MAE, y la raíz de la desviación cuadrática media, RMSE. El indicador MAE es la diferencia existente entre dos variables continuas, es decir, el error absoluto medio es un promedio de errores absolutos, dónde una de las variables es la predicción y la otra el valor verdadero, cuya fórmula es la siguiente:

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |P_i - P_i^*|$$

*Ecuación 1: Fórmula que representa el modo de calcular el MAE*

Donde:

- $P_i$: potencia real.
- $P_i^*$: potencia de predicción.
- N: Número de percentiles.

El indicador RMSE se calcula, tal como expresa la ecuación, como la raíz del error medio cuadrático (raíz del valor medio del cuadrado del error). El RMSE tiene las mismas unidades que los datos que se van a predecir con el modelo: por ejemplo, en la predicción de la producción, el indicador RMSE tiene unidades de producción de energía.

$$RMSE = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (P_i - P_i^*)^2}$$

*Ecuación 2: Fórmula que representa el modo de calcular el RMSE*

Donde:

- $P_i$: potencia real.
- $P_i^*$: potencia de predicción.
- N: Número de percentiles.

Asimismo, el $R^2$ (coeficiente de determinación) también se suele utilizar, debido a que su objetivo principal es predecir resultados futuros. Dicho coeficiente, decreta la calidad del modelo de predicción calculado para replicar los resultados, y la proporción de variación de los resultados obtenidos. $R^2$ adquiere valores comprendidos entre 0 y 1 (la predicción tendrá una mayor precisión cuando el valor se encuentre cercano al 1). Se expresa mediante la siguiente fórmula (aplicada para modelos determinísticos):

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \cdot \sigma_Y^2}$$

*Ecuación 3: Fórmula que representa el modo de calcular el $R^2$*

Donde:

- $\sigma_{xy}$ es la covarianza de (X,Y)
- $\sigma_x^2$ es la varianza de la variable X
- $\sigma_y^2$ es la varianza de la variable Y

## 9.2 Modelos de predicción probabilísticos

Como se ha comentado anteriormente, los modelos de predicción determinísticos permiten cuantificar la precisión del modelo realizado, propiedad que comparte con los modelos de predicción probabilística, a ello, añadir que estos últimos permiten determinar también la fiabilidad y la nitidez del modelo elaborado.

La fiabilidad cuantifica la semejanza existente entre la probabilidad de la predicción a priori y la frecuencia observada a posteriori. Una predicción probabilística es fiable si dicha predicción parece extraída de la misma distribución que los datos de muestreo. La nitidez se define como la capacidad de una predicción para "concentrar la información probabilística sobre los resultados futuros". A diferencia de la fiabilidad, es una propiedad únicamente enfocada en la predicción y su evaluación no involucra los datos utilizados para la elaboración del modelo; así se puede construir fácilmente una previsión arbitrariamente precisa. Por lo tanto, la maximización de la nitidez debe estar sujeta a la fiabilidad.

Para los modelos de predicción probabilísticos, además de los indicadores mencionados anteriormente (cuya finalidad es validar o determinar la precisión del modelo), otros de los indicadores que se utilizan cuando se realizan modelos de predicción probabilística son el CRPS (puntuación de probabilidad de clasificación continua) y el RMSD (raíz de la desviación cuadrática media).

Por un lado, el CRPS es una puntuación robusta que está diseñada de tal manera que mide tanto la fiabilidad como la nitidez. Una ventaja del CRPS es que reduce al error absoluto si el pronóstico es determinista, y esta puntuación, por lo tanto, permite la comparación entre probabilístico y previsiones de puntos. La fórmula que expresa el cálculo del CRPS es la siguiente:

$$CRPS\,(F,x) = \int_{-\infty}^{\infty} (F(y) - 1\{x \leq y\})^2 dy$$

***Ecuación 4: Fórmula que representa el modo de calcular el CRPS***

Donde:

- El símbolo “1”: representa la función escalón de Heaviside (es función discontinua cuyo valor es 0 para cualquier argumento negativo, y 1 para cualquier argumento positivo, incluido el cero).
- F: es la función de probabilidad acumulada.
- x: es el valor analizado para la variable.
- y: el valor previsto.

Cabe destacar que el CRPS al igual el RMSE, tiene las mismas unidades que los valores que se van a predecir, lo que permite una mejor interpretación de los resultados obtenidos. Asimismo, debido a su relación con el error

absoluto, el hecho de obtener un CRPS bajo indica un pronóstico probabilístico preciso.

Asimismo, la fórmula que permite calcular el RMSD es la enunciada a continuación:

$$RMSD = \sqrt{\frac{1}{N+1} \cdot \sum_{i=1}^{N+1} \left( S_i - \frac{M}{N+1} \right)^2}$$

*Ecuación 5: Fórmula que representa el modo de calcular el RMSD*

Donde

- N: Número de percentiles.
- $S_i$: es el número de casos en el segmento i.
- M: es el total de casos.

# 10. Modelos de predicción

## 10.1 Modelo climatológico

Es un modelo clásico en meteorología. Este método ofrece una técnica simple para predecir el clima porque se basa en tendencias pasadas. Los meteorólogos tienden a utilizarlo una vez han sido revisadas las estadísticas meteorológicas recopiladas a lo largo de varios años y de esta forma realizar el cálculo de los promedios. Predicen el clima para un día y ubicación específicos en función de las condiciones climáticas para ese mismo día durante varios años en el pasado.

El método de climatología solo funciona bien cuando el patrón climático es similar al esperado para la época del año elegida. Si el patrón es bastante inusual para la época del año, el método de climatología a menudo fallará.

Por tanto, la forma de realizar la predicción, se basa en adquirir el valor medio de observaciones o medidas en el pasado, con la intención de utilizarlos para la predicción de días con características similares. Por ejemplo, un meteorólogo puede dar una predicción probabilística de la temperatura media en septiembre en Logroño, analizando los valores medidos que hayan sido obtenidos en los últimos años. Es decir, obtiene un conjunto de valores medidos en el mes de septiembre en los últimos años y analizándolos estadísticamente proporciona una predicción probabilística en forma de percentiles (hay una probabilidad del 10% de que la temperatura media sea 0.7ºC, del 90% que se sitúe entre las 4.9 y los 15.5ºC, etc.).

Para este trabajo en cuestión, la predicción de la generación fotovoltaica en Alcolea del Río ha sido calculada, para cada hora del día, junto a la distribución estadística de lo generado en la misma hora en todos los datos del grupo de entrenamiento. Dicha distribución estadística ha sido calculada en forma de percentiles por medio de la hoja de cálculo Excel. Para aclarar cual ha sido el procedimiento llevado a cabo, se procede a explicarlo mediante un ejemplo.

En primer lugar, para realizar el análisis mediante la hoja de cálculo Excel, la muestra fue fraccionada en dos partes claramente diferenciadas:

- **Grupo de entrenamiento:** serán los datos utilizados para entrenar el modelo y optimizarlo, con la finalidad de ser posteriormente evaluado por los datos correspondientes al grupo de test y validar si se obtuvo

una buena predicción. Se utilizaron 16318 valores de la muestra, lo que corresponde a un total de 680 días de forma aproximada.

- **Grupo de test:** serán los datos utilizados para evaluar el modelo creado mediante el grupo de entrenamiento anteriormente descrito. Dicha muestra está formada por 3672 valores con una periodicidad de 1 hora, lo que corresponde a un total de 153 días aproximadamente.

Una vez se han elaborado ambas muestras, se utilizarán los datos pertenecientes al grupo de entrenamiento para realizar el modelo de predicción. Teniendo todos los datos pertinentes, se procede a ordenar la generación producida por la planta fotovoltaica para cada uno de los días desde las 0:00 horas hasta las 23:00 horas. La finalidad de realizar este paso es para poder obtener los percentiles correspondientes para cada día y en cada una de las horas del día (lógicamente para las horas nocturnas dichos percentiles serán 0 debido a que no hay ningún tipo de generación por parte de la planta fotovoltaica).

Por ejemplo, para obtener los percentiles correspondientes a las 12:00 horas, será necesario utilizar los valores de generación producidos a las 12:00 de los valores correspondientes al grupo de entrenamiento de cada uno de los días. Una vez realizado esto, se procede a utilizar una fórmula de la hoja de cálculo que permite obtener los diferentes percentiles (la fórmula utilizada es Percentil.INC, la cual devuelve el k-ésimo percentil de los valores en un rango, donde k está en el rango de 0 a 1, ambos incluidos) para poder realizar con ellos la distribución estadística que permita verificar la calidad de la predicción del modelo.

Asimismo, como los percentiles son una herramienta muy importante dentro de este trabajo, se procede a explicar de forma somera, qué son los percentiles. El percentil es una medida de posición usada en estudios estadísticos que indica, una vez ordenados los datos de menor a mayor, el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo de observaciones. Por ejemplo, el percentil 35 es el valor bajo el cual se encuentran el 35 por ciento de los resultados pertinentes.

Se representan con la letra P. Para el percentil i-ésimo, donde la i toma valores del 1 al 99. Generalmente, no se utilizan el percentil 0 y 100 porque pueden perjudicar a la predicción. Los percentiles principales son los siguientes

- $P_{25} = Q_1$.
- $P_{50} = Q_2$ = mediana.
- $P_{75} = Q_3$.

En la siguiente tabla, se presenta un ejemplo donde pueden apreciarse algunas de las horas de las cuales se han calculado dichos percentiles cuya finalidad es la de escoger los percentiles más significativos para con ellos representar la distribución estadística de dicho modelo.

| ***Percentil*** | **8:00** | **9:00** | **10:00** | **11:00** | **12:00** |
|---|---|---|---|---|---|
| ***0,05*** | 3 | 53 | 177,8 | 275,55 | 334,85 |
| ***0,10*** | 7 | 125 | 293,8 | 447,9 | 496 |
| ***0,15*** | 10 | 182,4 | 436,25 | 631,3 | 706,55 |
| ***0,20*** | 20 | 263 | 602,8 | 779,4 | 950,4 |
| ***0,25*** | 34,75 | 302 | 671,75 | 1000,25 | 1173,5 |
| ***0,30*** | 53 | 357,4 | 835 | 1137,7 | 1322,8 |
| ***0,35*** | 88,65 | 434,65 | 914,65 | 1224,65 | 1396 |
| ***0,40*** | 126,2 | 483,2 | 962,2 | 1270 | 1429,2 |
| ***0,45*** | 172 | 562,8 | 1000 | 1295,55 | 1447 |
| ***0,50*** | 236 | 666,5 | 1055 | 1322,5 | 1467,5 |
| ***0,55*** | 286,45 | 725,45 | 1090 | 1340 | 1486 |
| ***0,60*** | 333,8 | 756,4 | 1115,4 | 1361,4 | 1499,4 |
| ***0,65*** | 364,05 | 780 | 1141 | 1384,35 | 1519 |
| ***0,70*** | 394 | 804 | 1165 | 1400 | 1530,6 |
| ***0,75*** | 410 | 827 | 1183 | 1421,25 | 1552,25 |
| ***0,80*** | 440 | 842 | 1201,2 | 1446 | 1575 |
| ***0,85*** | 458,15 | 870 | 1230 | 1474,15 | 1602,15 |
| ***0,90*** | 480 | 910,6 | 1276,1 | 1525 | 1638 |
| ***0,95*** | 522,05 | 968,1 | 1329,05 | 1580,05 | 1720,15 |

***Tabla 2: Ejemplo de como se han calculado los percentiles para diferentes horas del día***

A continuación, se mostrará la distribución estadística del modelo climatológico, el cuál está únicamente realizado con los percentiles más significativos (Q05, Q25, Q50, Q75 y Q95). Cabe destacar que, dicha gráfica

permanecerá siempre igual, es decir, el modelo climatológico es un modelo persistente y, por tanto, los percentiles no van a cambiar, siempre van a ser los mismos, independientemente de la hora, del día etc.

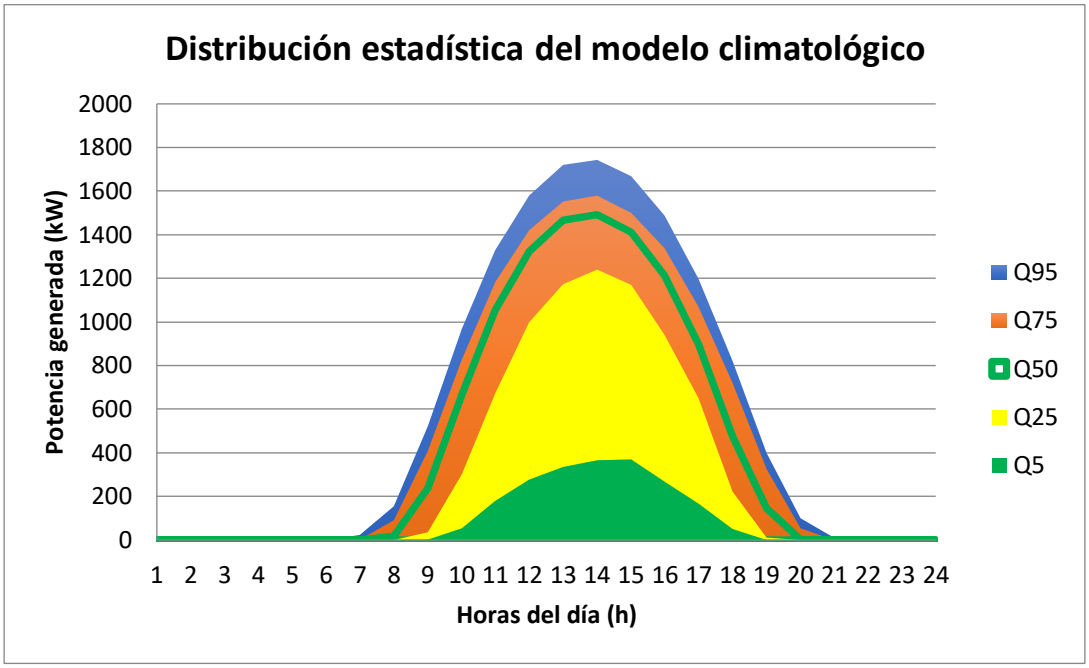


*Ilustración 10: Distribución probabilística del modelo climatológico*

Las siguientes tabla y gráfica, muestran la comparación existente en tres días consecutivos de los percentiles que constituyen el modelo climatológico (en los tres días permanecen igual, por lo comentado anteriormente) y la potencia producida real por parte de la planta fotovoltaica. En la gráfica se puede observar con bastante claridad, como se ajusta la potencia real producida al modelo climatológico, la cual nos proporciona información relevante, con ello se pretende explicar que, según la gráfica, el 95% de las ocasiones los valores de producción se encontrarán por debajo de los valores que muestran el percentil Q95. Quedando estos valores por encima de la mediana (los valores reales deberían ajustarse a la mediana, puesto que los errores se han realizado entorno a ella).

| Q95 | Q75 | Q50 | Q25 | Q05 | Real |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| 22 | 2 | 0 | 0 | 0 | 12 |
| 156 | 89 | 18 | 0 | 0 | 140 |
| 519,6 | 410 | 234 | 34 | 3 | 513 |
| 967,2 | 826 | 665 | 302 | 53 | 1052 |
| 1332 | 1183 | 1056 | 671 | 177,2 | 1356 |
| 1575,2 | 1421 | 1322 | 1001 | 274,2 | 1571 |
| 1720,6 | 1553 | 1468 | 1180 | 334,4 | 1730 |
| 1750 | 1579 | 1492 | 1240 | 362,8 | 1763 |
| 1670 | 1500 | 1411 | 1169 | 369,2 | 1698 |
| 1488,2 | 1337 | 1212 | 934 | 268,6 | 1503 |
| 1198,2 | 1068 | 897 | 651 | 165,8 | 1149 |
| 816,8 | 717 | 474 | 223 | 51 | 828 |
| 398,2 | 327 | 144 | 12 | 0 | 420 |
| 101,2 | 55 | 3 | 0 | 0 | 80 |
| 10 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | 2 | 0 | 0 | 0 | 8 |
| 156 | 89 | 18 | 0 | 0 | 138 |
| 519,6 | 410 | 234 | 34 | 3 | 542 |
| 967,2 | 826 | 665 | 302 | 53 | 1001 |
| 1332 | 1183 | 1056 | 671 | 177,2 | 1341 |
| 1575,2 | 1421 | 1322 | 1001 | 274,2 | 1558 |
| 1720,6 | 1553 | 1468 | 1180 | 334,4 | 1669 |
| 1750 | 1579 | 1492 | 1240 | 362,8 | 1694 |
| 1670 | 1500 | 1411 | 1169 | 369,2 | 1667 |
| 1488,2 | 1337 | 1212 | 934 | 268,6 | 1246 |
| 1198,2 | 1068 | 897 | 651 | 165,8 | 612 |
| 816,8 | 717 | 474 | 223 | 51 | 444 |
| 398,2 | 327 | 144 | 12 | 0 | 101 |
| 101,2 | 55 | 3 | 0 | 0 | 26 |
| 10 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | 2 | 0 | 0 | 0 | 10 |
| 156 | 89 | 18 | 0 | 0 | 174 |
| 519,6 | 410 | 234 | 34 | 3 | 598 |

| | | | | | |
|---|---|---|---|---|---|
| 967,2 | 826 | 665 | 302 | 53 | 934 |
| 1332 | 1183 | 1056 | 671 | 177,2 | 1180 |
| 1575,2 | 1421 | 1322 | 1001 | 274,2 | 1558 |
| 1720,6 | 1553 | 1468 | 1180 | 334,4 | 1521 |
| 1750 | 1579 | 1492 | 1240 | 362,8 | 1684 |
| 1670 | 1500 | 1411 | 1169 | 369,2 | 1619 |
| 1488,2 | 1337 | 1212 | 934 | 268,6 | 1435 |
| 1198,2 | 1068 | 897 | 651 | 165,8 | 1161 |
| 816,8 | 717 | 474 | 223 | 51 | 863 |
| 398,2 | 327 | 144 | 12 | 0 | 183 |
| 101,2 | 55 | 3 | 0 | 0 | 32 |
| 10 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |

*Tabla 3: Representación de los percentiles de tres días consecutivos junto con la producción real, modelo climatológico*

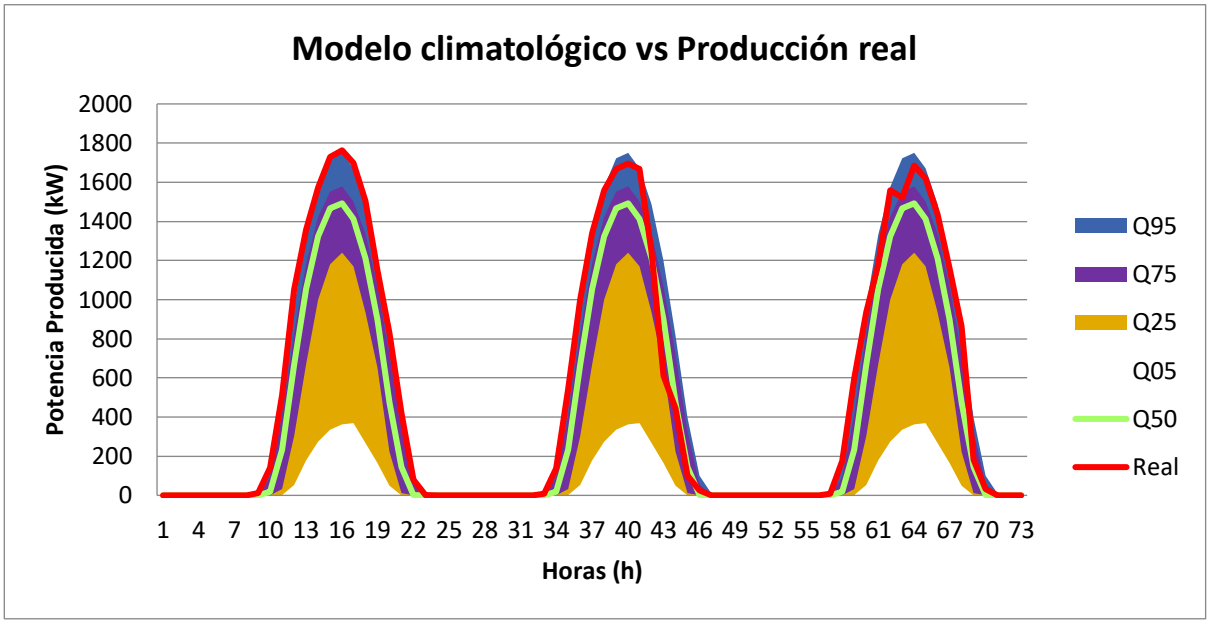


*Ilustración 11: Representación de tres días consecutivos del modelo climatológico frente a la producción real*

### 10.1.1 Código para el análisis mediante la herramienta Rstudio

Código utilizado para el análisis mediante Rstudio, para la obtención de los criterios evaluadores del modelo de predicción climatológico:

```
# Programa para realizar el cálculo de los criterios de evaluación directamente con los mismos datos
# que se usaron en las predicciones determinísticas
options(java.parameters = "- Xmx2048m")
```

```
setwd("N:/modelos")

# -------

library(scoringRules)
library(readxl)
library(matrixStats)
library(stats)
library(KScorrect)
library(xlsx)
library(SpecsVerification)
library(quantreg)

rsq <- function (x, y) cor(x, y) ^ 2
Datos <- read_excel("Percentiles_clima.xlsx")

df <- data.frame(Datos)
entrenamiento <- as.matrix(df[1:16318, 1:19])
pot_ent <- df[1:16318, 20]
pot_ent <- as.numeric(pot_ent)
testing <- as.matrix(df[16319:19990, 1:19])
pot_test <- df[16319:19990, 20]
pot_test <- as.numeric(pot_test)

crps_ent <- matrix(nrow=16318, ncol=1, NA)
crps_test <- matrix(nrow=3672, ncol=1, NA)

crps_ent <- crps_sample(pot_ent, entrenamiento, method = "edf", w = NULL,
bw = NULL,num_int = FALSE, show_messages = TRUE)
crps_test <- crps_sample(pot_test, testing, method = "edf", w = NULL, bw =
NULL,num_int = FALSE, show_messages = TRUE)

crps_entrenamiento <- mean(crps_ent)
crps_testing <- mean(crps_test)

rmse_entrenamiento <- sqrt(mean((pot_ent-entrenamiento[,10])^2))
mae_entrenamiento <- mean(abs(pot_ent-entrenamiento[,10]))
rsq_entrenamiento <- rsq(pot_ent, entrenamiento[,10])

rmse_testing <- sqrt(mean((pot_test-testing[,10])^2))
mae_testing <- mean(abs(pot_test-testing[,10]))
rsq_testing <- rsq(pot_test,testing[,10])

rank.hist <- Rankhist(entrenamiento, pot_ent)
PlotRankhist(rank.hist, mode="raw")
abline(h=16318/20, col="red",lwd=3)
teo <- matrix(nrow=1, ncol=20, 16318/20)
rmsd_entrenamiento <- sqrt(1/20*sum((rank.hist-teo)^2))

rank.hist <- Rankhist(testing, pot_test)
```

```
PlotRankhist(rank.hist, mode="raw")
abline(h=3672/20, col="red",lwd=3)
teo <- matrix(nrow=1, ncol=20, 3672/20)
rmsd_testing <- sqrt(1/20*sum((rank.hist-teo)^2))

sink(paste0("Modelo_climatologico.txt"))
print("\n")
print("\n")
print(paste0("CRPS train: ", crps_entrenamiento))
print(paste0("RMSE train: ", rmse_entrenamiento))
print(paste0("MAE train: ", mae_entrenamiento))
print(paste0("R2 train: ", rsq_entrenamiento))
print(paste0("RMSD train: ", rmsd_entrenamiento))
print(paste0("CRPS test: ", crps_testing))
print(paste0("RMSE test: ", rmse_testing))
print(paste0("MAE test: ", mae_testing))
print(paste0("R2 test: ", rsq_testing))
print(paste0("RMSD test: ", rmsd_testing))
sink()
```

Tras ejecutar el código los resultados obtenidos de los percentiles de los datos de test han sido los siguientes:

| | CRPS | RMSE | MAE | $R^2$ | RMSD |
|---|---|---|---|---|---|
| ***Entrenamiento*** | 83,19345 | 248,83940 | 116,9331 | 0,8185 | 23,4391 |
| ***Test*** | 60,64788 | 164,24777 | 81,0676 | 0,9169 | 51,5445 |

**Tabla 4: Resultados obtenidos para los grupos de entrenamiento y test, después de ejecutar Rstudio**

Se ha realizado la compilación del código para ambos grupos, puede apreciarse como los criterios evaluadores de los modelos de predicción comentados anteriormente (CRPS, RMSE, MAE…), son mejores para el grupo de test (datos que se han utilizado para evaluar el modelo de predicción elaborado) que para los datos del grupo de entrenamiento con los cuales se ha construido dicho modelo.

Además, dicho código ha proporcionado los histogramas de ambos grupos, mostrando como se ajustan los valores en cada caso, ajustándose mejor en este caso a los datos del grupo de entrenamiento, difiriendo los errores calculados anteriormente. Se procede a dar una explicación breve de que es un histograma para poder aclarar en la medida de lo posible lo que muestran realmente.

Un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia (en este caso, la frecuencia de las barras son los percentiles, por lo que tienen una frecuencia del 5%) de los valores representados. Sirven para obtener una primera visión general de la distribución la muestra, respecto a una característica, cuantitativa y continua (para este caso en particular será el número de ocasiones en los que la potencia producida se encuentra dentro dicho baremo). De esta manera ofrece una visión de grupo permitiendo observar una preferencia, o tendencia, por parte de la muestra.

En este primer histograma, que corresponde a los datos del grupo de entrenamiento, la línea roja determina el número de casos (en este caso sería un total de 820 observaciones aproximadamente) que no deberían sobrepasar ninguna de las barras, situación que se cumple en este caso, proporcionando homogeneidad a la muestra.

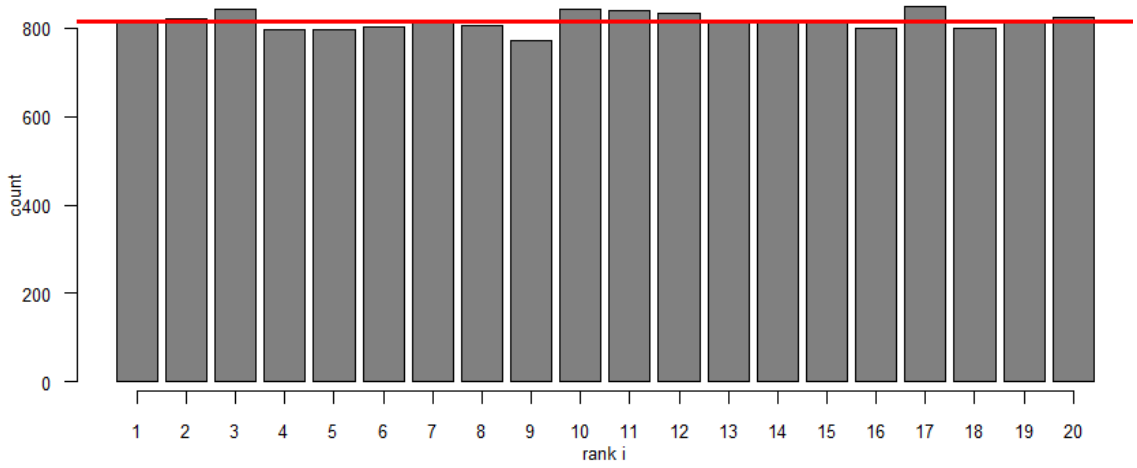


*Ilustración 12: Histograma perteneciente al grupo de entrenamiento, modelo climatológico*

En este segundo histograma, que corresponde a los datos del grupo de test, la homogeneidad que se mostraba en el histograma anterior, de los datos del grupo de entrenamiento, no se produce.

Esta situación está íntimamente relacionada con la gráfica expuesta anteriormente del modelo climatológico frente a la producción real, donde la producción real se posicionaba generalmente en percentiles altos. Además, la línea roja que demarca el número de casos (en el caso de este histograma el número de observaciones que debería haber se sitúa entorno a 180 aproximadamente) que no deben sobrepasar dicha marca, es sobrevalorada en

los percentiles altos e infravalorada en los percentiles bajos, por tanto, no se cumple en algunas de las barras y esto puede producirse por una sobrevaloración de los datos de la muestra o por los meses recogidos en dicha muestra, los cuales tengan realmente una mayor producción que la que realmente se han predicho.

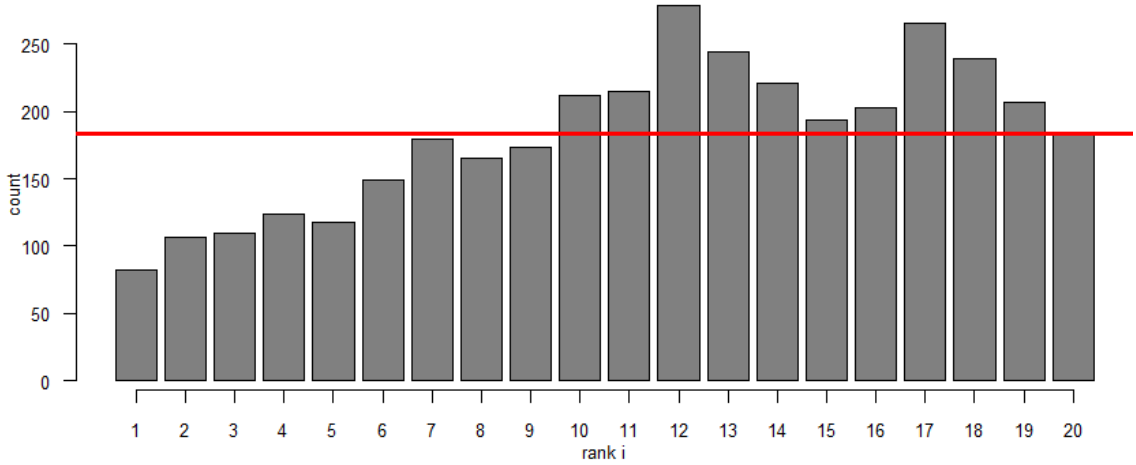


*Ilustración 13: Histograma perteneciente al grupo de test, modelo climatológico*

## 10.2 Modelo persistente probabilístico

Se podría considerar como el modelo de referencia en predicciones determinísticas. Dicho modelo, supone que el valor de predicción es el último valor conocido. Por ejemplo, el valor previsto con este modelo para lo que se generará mañana en la planta fotovoltaica a las 12 horas será igual al último valor conocido de generación a esa hora (el de hoy a las 12 si se conociera ya, si no el de ayer a las 12).

Dentro del probabilísmo no existe el persistente tal cual lo hemos definido antes, ya que con un único valor (el esperado) es imposible obtener distribuciones probabilísticas. En este caso, se ha optado por llamar modelo persistente probabilístico al que proporciona la distribución estadística, en forma de percentiles, de la generación en el instante futuro, obtenida a partir de los valores medidos de generación en los últimos días a la misma hora. La cuestión es determinar cuantos días escoger para realizar el cálculo de dicha distribución (no sería lo mismo tomar 15 valores que tomar 25, por ejemplo).

A continuación, se explicará el procedimiento llevado a cabo para la selección de la ventana fija de días atrás que genere mejores resultados en los datos pertenecientes al grupo de entrenamiento para posteriormente aplicar dicha venta sobre los resultados de los valores del grupo de test y así poder evaluar las predicciones del modelo. Asimismo, se explicará el procedimiento de análisis y de obtención de resultados para dos modelos, modelo de día similar (hoy – hoy) y modelo de día similar (hoy – mañana).

### 10.2.1 Modelo día similar (hoy – hoy)

En primer lugar, realizar un matiz que es el fundamental para diferenciar a este modelo con el modelo de predicción de “hoy-mañana”. Con este modelo se puede realizar una predicción probabilística siempre y cuando se tengan los datos suficientes del día de hoy para poder realizar dichas predicciones, en caso de no ser así, se deberá utilizar el modelo de “hoy-mañana”. Un ejemplo sencillo que explique esta diferencia es el siguiente: si para el día de hoy a las 11 se conocieran los datos necesarios para la predicción se utilizaría el modelo de predicción de “hoy-hoy”, en caso de no disponer de dichos datos, se deberá usar el de ayer a las 11.

Para realizar el análisis mediante el cual se elegirá la ventana fija más apropiada, a través de la cual determinar el número de días idóneo para realizar

las predicciones, la muestra fue fraccionada en dos partes claramente diferenciadas:

- **Grupo de entrenamiento:** serán los datos utilizados para entrenar el modelo y optimizarlo, con la finalidad de ser posteriormente evaluado por los datos correspondientes al grupo de test y validar si se obtuvo una buena predicción. Se utilizaron 16318 valores de la muestra, lo que corresponde a un total de 680 días de forma aproximada.

- **Grupo de test:** serán los datos utilizados para evaluar el modelo creado mediante el grupo de entrenamiento anteriormente descrito. Dicha muestra está formada por 3672 valores con una periodicidad de 1 hora, lo que corresponde a un total de 153 días aproximadamente.

Una vez se han elaborado ambas muestras, se utilizarán los datos pertenecientes al grupo de entrenamiento para realizar el modelo de predicción.

Para elegir la ventana fija (el número de días, días atrás) apropiada para poder obtener los mejores resultados posibles, se deberá analizar todas las horas de los datos del grupo de entrenamiento y, una vez se haya conseguido la predicción probabilística para ellas con distinto número de días, se seleccionará en función de los criterios evaluadores (CRPS, MAE, RMSE…) cuál es el número de días más adecuado para posteriormente aplicarlo sobre los datos del grupo de test y así obtener las predicciones probabilísticas pertinentes. Dicho procedimiento de selección, se llevará a cabo mediante la ejecución de un código realizado en Rstudio.

#### 10.2.1.1 Código para el análisis mediante la herramienta Rstudio

Código utilizado para el análisis mediante Rstudio, para la obtención del número de días que mejores resultados proporcionará (compromiso entre el menor CRPS y el menor RMSD):

```
# Programa para realizar el modelo persistente directamente con los mismos datos
# que se usaron en las predicciones determinísticas

#options(java.parameters = "- Xmx2048m")

#setwd("N:/modelos")
```

```
# -------

library(scoringRules)
library(readxl)
library(matrixStats)
library(stats)
library(KScorrect)
library(xlsx)
library(SpecsVerification)
library(quantreg)

rsq <- function (x, y) cor(x, y) ^ 2

Datos <- read_excel("Datos.xlsx")

resultados <- matrix(nrow=34, ncol=5, NA)
colnames(resultados) <- c("CRPS","RMSE", "MAE", "R2", "RMSD")
crps_dia <- matrix(nrow=680, ncol=1, NA)
cuantiles_totales <- matrix(nrow=16320, ncol=19, NA)
prediccion <- matrix(nrow=16320, ncol=1, NA)

# Empiezo con el primer día de entrenamiento
for (lim_inf in c(3:36)){
        conta <- 1
        for (ind in seq(864, 16296, by=24 )){
                i <- ind
                gene <- Datos$ene[(ind):(ind+23)]
                ensemble <- matrix(nrow=24, ncol=36, NA)
                cuantiles <- matrix(nrow=24, ncol=19, NA)
                crps_valores <- matrix(nrow=24, ncol=1, NA)

# Creo ensemble para el dia D. Obtengo las predicciones a base de los cuantiles
con la ventana escogida
                for (j in c(1:24)){
# Relleno la fila de la hora correspondiente del ensemble
                        for (d in c(1:36)){
                                k <- i+j-1-d*24
                                if (k>=1) {
                                        ensemble[j,d] <- Datos$ene[k]
                                } else {
                                        ensemble[j,d] <- NA
                                }
                        }
                }
```

```
            for (j in c(1:24)){
                cuantiles[j,] <- quantile(ensemble[j,1:lim_inf], probs = seq(0.05, 0.95, 0.05), na.rm = TRUE, names = TRUE, type = 6)
                crps_valores[j] <- crps_sample(gene[j], cuantiles[j], method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)
            }
            crps_dia[conta] <- mean(crps_valores)
            print(paste0(conta, "     ", lim_inf, "   ", crps_dia[conta]))
            prediccion[(ind):(ind+23)] <- cuantiles[ ,10]
            cuantiles_totales[(ind):(ind+23),] <- cuantiles
            conta <- conta +1
        }


# Evaluo el crps, rmse, mae, mre, rix para todo el grupo de entrenamiento
        crps_medio <- mean(crps_dia, na.rm=TRUE)
        rmse <- sqrt(mean((Datos$ene[864:16319]-prediccion[864:16319])^2))
        mae <- mean(abs(Datos$ene[864:16319]-prediccion[864:16319]))
        rsq_testing <- rsq(prediccion[864:16319],Datos$ene[864:16319])

        rank.hist <- Rankhist(as.matrix(cuantiles_totales[864:16319,]), as.matrix(Datos$ene[864:16319]))
        teo <- matrix(nrow=1, ncol=20, 15456/20)
        rmsd_total <- sqrt(1/20*sum((rank.hist-teo)^2))
        print(paste0(crps_medio, "  ", rmse, "   ", mae, "  ", rsq_testing))
        resultados[lim_inf-2,1] <- crps_medio
        resultados[lim_inf-2,2] <- rmse
        resultados[lim_inf-2,3] <- mae
        resultados[lim_inf-2,4] <- rsq_testing
        resultados[lim_inf-2,5] <- rmsd_total
}

file_e <- paste0("Resultados_ventana_crps_solodiasatras_ent.xlsx")
write.xlsx(resultados, file=file_e, sheetName = "Resultados", col.names = TRUE, row.names = FALSE, append = FALSE)
```

Después de ejecutar el código se han obtenido los resultados necesarios los cuales se presentarán a continuación, determinando cuál es el número de días más adecuado para obtener las predicciones probabilísticas.

| Nº días | CRPS | RMSE | MAE | R2 | RMSD |
|---|---|---|---|---|---|
| **3** | 119,0047 | 223,8242 | 86,91353 | 0,84245 | 602,6216 |
| **4** | 134,2361 | 211,6746 | 84,86208 | 0,857417 | 477,5848 |
| **5** | 147,6491 | 215,2375 | 84,04755 | 0,855231 | 403,9873 |
| **6** | 160,2174 | 210,9737 | 83,82039 | 0,860174 | 347,6114 |
| **7** | 171,6734 | 215,1924 | 84,78976 | 0,85602 | 298,1234 |
| **8** | 182,5763 | 213,27 | 84,75856 | 0,858327 | 261,8909 |
| **9** | 192,4194 | 214,9307 | 85,05626 | 0,857218 | 241,5857 |
| **10** | 200,0828 | 214,0315 | 85,39653 | 0,858333 | 210,1893 |
| **11** | 206,1844 | 215,3313 | 85,46522 | 0,857578 | 179,0845 |
| **12** | 212,4537 | 214,3048 | 85,45232 | 0,858883 | 169,4891 |
| **13** | 218,5643 | 215,6518 | 86,04927 | 0,857564 | 149,1216 |
| **14** | 224,4024 | 214,9355 | 86,20586 | 0,8584 | 135,203 |
| **15** | 229,7505 | 216,5894 | 86,81535 | 0,856606 | 120,3659 |
| **16** | 242,9509 | 217,0318 | 87,61416 | 0,856416 | 100,7465 |
| **17** | 238,9886 | 216,4911 | 87,12037 | 0,857136 | 111,3883 |
| **18** | 234,4571 | 215,6717 | 86,80301 | 0,85778 | 113,9674 |
| **19** | 246,937 | 218,3682 | 88,16013 | 0,855045 | 102,8594 |
| **20** | 247,2765 | 218,3106 | 88,5219 | 0,855105 | 111,7124 |
| **21** | 247,4937 | 219,5104 | 88,98389 | 0,853909 | 109,1886 |
| **22** | 247,3347 | 219,8456 | 89,20474 | 0,853464 | 106,8066 |
| **23** | 247,1361 | 220,5606 | 89,55273 | 0,852899 | 114,5516 |
| **124** | 246,9237 | 220,8708 | 89,94648 | 0,852421 | 104,1689 |
| **25** | 246,663 | 221,6761 | 90,38428 | 0,851593 | 116,351 |
| **26** | 246,0469 | 222,0128 | 90,63272 | 0,85112 | 126,6032 |
| **27** | 245,6949 | 222,7686 | 90,80826 | 0,850393 | 121,676 |
| **28** | 245,1505 | 222,8343 | 91,06829 | 0,850332 | 114,7831 |
| **29** | 244,4006 | 223,1239 | 91,30283 | 0,850179 | 124,2592 |
| **30** | 243,8523 | 223,1483 | 91,32576 | 0,850094 | 131,1642 |
| **31** | 243,1729 | 223,9955 | 91,58172 | 0,849103 | 137,7863 |
| **32** | 242,4923 | 223,9685 | 91,64978 | 0,849257 | 131,2203 |
| **33** | 241,9407 | 224,7227 | 91,9209 | 0,848548 | 127,3352 |
| **34** | 241,3688 | 225,0993 | 92,1754 | 0,848077 | 130,8841 |
| **35** | 240,88 | 225,468 | 92,32942 | 0,847744 | 133,5012 |
| **36** | 240,3002 | 225,3053 | 92,3093 | 0,848078 | 141,8455 |

*Tabla 5: Elección ventana fija óptima, modelo persistente probabilístico "Hoy-hoy"*

Tras observar los datos, se determina que el día que mejor compromiso obtiene entre el menor CRPS y menor RMSD, es el de 16 días atrás. Una vez se ha elegido el número de días que genera las mejores predicciones, se procede a aplicar dicha ventana sobre los datos del grupo de test para obtener los resultados de la predicción probabilística.

```
# Se ha comprobado que lo mejor es 16 días atrás, lo recompongo con 16 para sacar histograma

lim_inf <- 16
conta <- 1
for (ind in seq(2, 16322, by=24 )){
        i <- ind
        gene <- Datos$ene[(ind):(ind+23)]
        ensemble <- matrix(nrow=24, ncol=36, NA)
        cuantiles <- matrix(nrow=24, ncol=19, NA)
        crps_valores <- matrix(nrow=24, ncol=1, NA)
# Creo ensemble para el día D. Obtengo las predicciones a base de los cuantiles con la ventana escogida
        for (j in c(1:24)){
# Relleno la fila de la hora correspondiente del ensemble
                for (d in c(1:36)){
                        k <- i+j-1-d*24
                        if (k>=1) {
                                ensemble[j,d] <- Datos$ene[k]
                        } else {
                                ensemble[j,d] <- NA
                        }
                }
        }
        for (j in c (1:24)) {
                cuantiles[j,] <- quantile(ensemble[j,1:lim_inf], probs = seq(0.05, 0.95, 0.05), na.rm = TRUE, names = TRUE, type = 6)
crps_valores[j] <- crps_sample(gene[j], cuantiles[j], method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)
        }
        crps_dia[conta] <- mean(crps_valores)
        print(paste0(conta, "     ", lim_inf, "   ", crps_dia[conta]))
        prediccion[(ind-2):(ind-2+23)] <- cuantiles[ ,10]
        cuantiles_totales[(ind-2):(ind-2+23),] <- cuantiles
        conta <- conta +1
}
        # Evaluo el crps, rmse, mae, mre, rix para todo el grupo de test
rank.hist <- Rankhist(cuantiles_totales, Datos$ene[2:16322])
teo <- matrix(nrow=1, ncol=20, 16320/20)
PlotRankhist(rank.hist, mode="raw")
abline(h=16320/20, col="red",lwd=3)

file_e <- paste0("Cuantiles_persistente_16dias_ent.xlsx")
```

```
write.xlsx(cuantiles_totales, file=file_e, sheetName = "Percentiles_test", col.names = TRUE, row.names = FALSE, append = FALSE)
```

Una vez obtenidos los percentiles de la distribución probabilista, se procede a obtener los criterios evaluadores generados por los datos del grupo de test.

```
# Programa para obtener los criterios evaluadores

options(java.parameters = "- Xmx2048m")

# -------

library(scoringRules)
library(readxl)
library(matrixStats)
library(stats)
library(KScorrect)
library(xlsx)
library(SpecsVerification)
library(quantreg)

rsq <- function (x, y) cor(x, y) ^ 2

Datos <- read_excel("Cuantiles_persistente_16dias.xlsx")

df <- data.frame(Datos)
testing <- as.matrix(df[1:3672, 1:19])
pot_test <- df[1:3672, 20]
pot_test <- as.numeric(pot_test)

crps_test <- matrix(nrow=3672, ncol=1, NA)

crps_test <- crps_sample(pot_test, testing, method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)

crps_testing <- mean(crps_test)

rmse_testing <- sqrt(mean((pot_test-testing[,10])^2))
mae_testing <- mean(abs(pot_test-testing[,10]))
rsq_testing <- rsq(pot_test,testing[,10])

rank.hist <- Rankhist(testing, pot_test)
```

```
PlotRankhist(rank.hist, mode="raw")
abline(h=3672/20, col="red",lwd=3)
teo <- matrix(nrow=1, ncol=20, 3672/20)
rmsd_testing <- sqrt(1/20*sum((rank.hist-teo)^2))

sink(paste0("Modelo_persistente_16_test.txt"))
print("\n")
print("\n")
print(paste0("CRPS test: ", crps_testing))
print(paste0("RMSE test: ", rmse_testing))
print(paste0("MAE test: ", mae_testing))
print(paste0("R2 test: ", rsq_testing))
print(paste0("RMSD test: ", rmsd_testing))
sink()
```

Los datos que se han conseguido para la ventana fija de 16 días aplicados sobre los datos del grupo de test se recogen en la siguiente tabla:

| | CRPS | RMSE | MAE | $R^2$ | RMSD |
|---|---|---|---|---|---|
| **Entrenamiento** | 160,2174 | 210,9737 | 83,8204 | 0,8602 | 127,6114 |
| **Test** | 45,54270 | 154,10031 | 59,4432 | 0,92837 | 28,39084 |

*Tabla 6: Resultados de los grupos de entrenamiento y test, modelo persistente probabilístico "Hoy-hoy"*

En el siguiente histograma, que corresponde a los datos del grupo de test, la línea roja determina el número de casos (en este caso sería un total de 180 observaciones aproximadamente) los cuales no deberían ser sobrepasados por ninguna de las barras siempre que se realice una predicción precisa y fiable, situación que se cumple parcialmente, proporcionando cierta homogeneidad a la muestra.

En este caso, ocurre lo contrario a lo que sucedía en el modelo climatológico y, es que ahora los valores se han infravalorado, es decir, los datos de la muestra reflejen una menor producción de la que realmente se ha predicho.

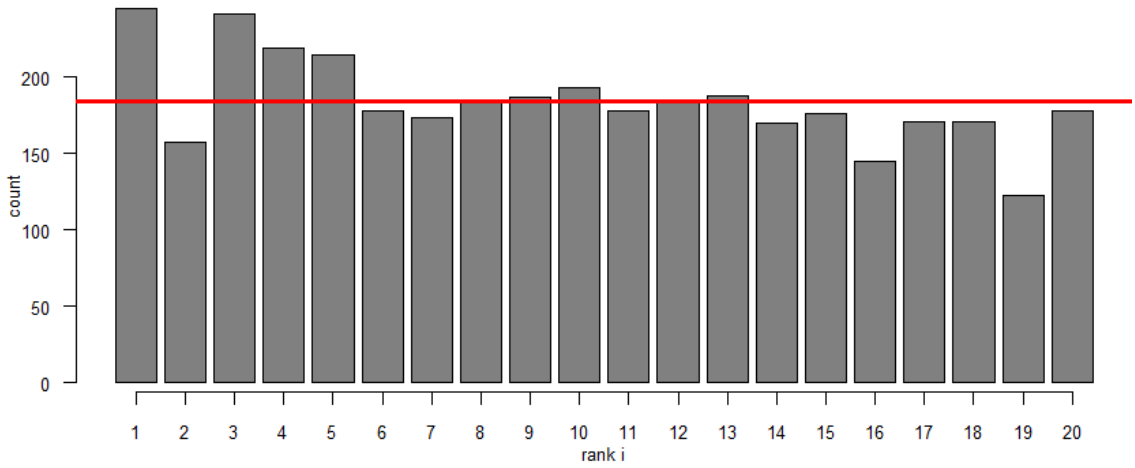


*Ilustración 14: Histograma grupo de test, modelo persistente probabilístico "Hoy-hoy"*

### 10.2.2 Modelo día similar (hoy – mañana)

El siguiente modelo, es similar al explicado anteriormente con la diferencia de que las predicciones probabilísticas se realizarán con un horizonte de predicción diferente, en vez de hacer predicciones de hoy para hoy se realizarán predicciones de hoy para mañana (matiz que se explicó anteriormente). La forma de proceder de este modelo es idéntica a la expuesta anteriormente.

En primer lugar, para realizar el análisis mediante el cual se elegirá el número de días idóneo para realizar las predicciones, la muestra será fraccionada en dos partes claramente diferenciadas, las cuales ya se expusieron para el modelo anterior.

Una vez elaboradas ambas muestras, se utilizarán los datos pertenecientes al grupo de entrenamiento para realizar el modelo de predicción.

Para elegir la ventana fija (el número de días, días atrás) apropiada para poder obtener los mejores resultados posibles, se deberá analizar todas las horas de los datos del grupo de entrenamiento, aunque en este caso para realizar la predicción se partirá del día D-1, debido a que este modelo parte del día anterior para obtener la información necesaria para realizar las predicciones.

Una vez se haya conseguido la predicción probabilística con distinto número de días, se seleccionará en función de los criterios evaluadores (CRPS, MAE, RMSE…) cuál es el número de días más adecuado para posteriormente aplicarlo sobre los datos del grupo de test y así obtener las predicciones probabilísticas pertinentes. Dicho procedimiento de selección, se llevará a cabo mediante la ejecución de un código realizado en Rstudio.

#### 10.2.2.1 Código para el análisis mediante la herramienta Rstudio

Código utilizado (cambia ligeramente respecto del anterior) para el análisis mediante Rstudio, para la obtención del número de días que mejores resultados proporcionará (compromiso entre el menor CRPS y el menor RMSD):

```
# Programa para realizar el modelo persistente directamente con los mismos
# datos que se usaron en las predicciones determinísticas

#options(java.parameters = "- Xmx2048m")

#setwd("N:/modelos")

# -------

library(scoringRules)
library(readxl)
library(matrixStats)
library(stats)
library(KScorrect)
library(xlsx)
library(SpecsVerification)
library(quantreg)

rsq <- function (x, y) cor(x, y) ^ 2

Datos <- read_excel("Datos.xlsx")

resultados <- matrix(nrow=34, ncol=5, NA)
colnames(resultados) <- c("CRPS","RMSE", "MAE", "R2", "RMSD")
crps_dia <- matrix(nrow=680, ncol=1, NA)
cuantiles_totales <- matrix(nrow=16320, ncol=19, NA)
prediccion <- matrix(nrow=16320, ncol=1, NA)

# Empiezo con el primer día de test
for (lim_inf in c(3:36)){
        conta <- 1
        for (ind in seq(864, 16296, by=24 )){
                i <- ind
                gene <- Datos$ene[(ind):(ind+23)]
                ensemble <- matrix(nrow=24, ncol=36, NA)
                cuantiles <- matrix(nrow=24, ncol=19, NA)
```

```
                crps_valores <- matrix(nrow=24, ncol=1, NA)
# Creo ensemble para el dia D. Obtengo las predicciones a base de los cuantiles con la ventana escogida
                for (j in c(1:24)){
# Relleno la fila de la hora correspondiente del ensemble
                        for (d in c(2:36)){
                                k <- i+j-1-d*24
                                if (k>=1) {
                                        ensemble[j,d] <- Datos$ene[k]
                                } else {
                                        ensemble[j,d] <- NA
                                }
                        }
                }
                for (j in c(1:24)){
                        cuantiles[j,] <- quantile(ensemble[j,1:lim_inf], probs = seq(0.05, 0.95, 0.05), na.rm = TRUE, names = TRUE, type = 6)
                        crps_valores[j] <- crps_sample(gene[j], cuantiles[j], method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)
                }
                crps_dia[conta] <- mean(crps_valores)
                print(paste0(conta, "     ", lim_inf, "   ", crps_dia[conta]))
                prediccion[(ind):(ind+23)] <- cuantiles[ ,10]
                cuantiles_totales[(ind):(ind+23),] <- cuantiles
                conta <- conta +1
        }
        # Evaluo el crps, rmse, mae, mre, rix para todo el grupo de test
        crps_medio <- mean(crps_dia, na.rm=TRUE)
        rmse <- sqrt(mean((Datos$ene[864:16319]-prediccion[864:16319])^2))
        mae <- mean(abs(Datos$ene[864:16319]-prediccion[864:16319]))
        rsq_testing <- rsq(prediccion[864:16319],Datos$ene[864:16319])

        rank.hist <- Rankhist(as.matrix(cuantiles_totales[864:16319,]), as.matrix(Datos$ene[864:16319]))
        teo <- matrix(nrow=1, ncol=20, 15456/20)
        rmsd_total <- sqrt(1/20*sum((rank.hist-teo)^2))
        print(paste0(crps_medio, "  ", rmse, "   ", mae, "  ", rsq_testing))
        resultados[lim_inf-2,1] <- crps_medio
        resultados[lim_inf-2,2] <- rmse
        resultados[lim_inf-2,3] <- mae
        resultados[lim_inf-2,4] <- rsq_testing
        resultados[lim_inf-2,5] <- rmsd_total
}
```

```
file_e <- paste0("Resultados_ventana_crps_solodiasatras_ent_mañana.xlsx")
write.xlsx(resultados, file=file_e, sheetName = "Resultados", col.names = TRUE,
row.names = FALSE, append = FALSE)
```

Después de ejecutar el código se han obtenido los resultados necesarios los cuales se presentarán a continuación, determinando cuál es el número de días más adecuado para obtener las predicciones probabilísticas.

| Nº días | CRPS | RMSE | MAE | R2 | RMSD |
|---|---|---|---|---|---|
| **3** | 115,5695 | 232,621 | 96,76113 | 0,826779 | 846,6113 |
| **4** | 129,7655 | 236,7301 | 94,64677 | 0,824607 | 654,9095 |
| **5** | 143,1266 | 222,8329 | 91,26525 | 0,842592 | 546,6002 |
| **6** | 156,072 | 225,3127 | 90,28073 | 0,841863 | 467,0025 |
| **7** | 167,7043 | 221,5861 | 89,90263 | 0,846256 | 400,1121 |
| **8** | 178,649 | 224,6324 | 90,11559 | 0,843546 | 356,445 |
| **9** | 188,846 | 220,3462 | 89,18613 | 0,849078 | 309,7113 |
| **10** | 196,5405 | 221,6924 | 89,26168 | 0,848376 | 282,8409 |
| **11** | 202,6536 | 219,8877 | 88,92665 | 0,850703 | 240,1938 |
| **12** | 209,4038 | 220,7084 | 88,78316 | 0,850559 | 217,4324 |
| **13** | 215,8542 | 219,4222 | 88,66126 | 0,852223 | 202,2411 |
| **14** | 221,7827 | 220,5375 | 89,16566 | 0,851179 | 180,7342 |
| **15** | 227,1415 | 219,7049 | 89,19919 | 0,852175 | 175,1826 |
| **16** | 231,7265 | 220,8221 | 89,71875 | 0,851066 | 152,7012 |
| **17** | 236,2841 | 219,9593 | 89,65297 | 0,852178 | 149,8588 |
| **18** | 240,4921 | 221,544 | 90,14088 | 0,850519 | 140,5004 |
| **19** | 244,5941 | 221,3468 | 90,3213 | 0,850766 | 125,6251 |
| **20** | 248,2057 | 222,629 | 90,85242 | 0,849452 | 136,6333 |
| **21** | 248,4487 | 222,5671 | 91,01556 | 0,849527 | 131,4924 |
| **22** | 248,3449 | 223,9995 | 91,47237 | 0,848013 | 120,3352 |
| **23** | 248,3095 | 223,8603 | 91,74495 | 0,848206 | 125,1306 |
| **124** | 248,2048 | 224,5192 | 92,03261 | 0,847722 | 134,603 |
| **25** | 247,9201 | 224,6629 | 92,29359 | 0,847468 | 126,873 |
| **26** | 247,2829 | 225,6176 | 92,61365 | 0,846438 | 133,2976 |
| **27** | 247,0482 | 225,6348 | 92,67192 | 0,846405 | 137,5815 |
| **28** | 246,4747 | 225,9475 | 92,86306 | 0,84628 | 135,6203 |
| **29** | 245,7562 | 225,6838 | 92,89389 | 0,846674 | 136,9677 |
| **30** | 245,2111 | 226,054 | 92,97939 | 0,846423 | 142,8018 |
| **31** | 244,5197 | 226,2725 | 93,06515 | 0,846086 | 144,0519 |
| **32** | 243,792 | 226,8531 | 93,40159 | 0,845452 | 151,9778 |
| **33** | 243,2002 | 226,9622 | 93,49518 | 0,845429 | 149,5101 |
| **34** | 242,6554 | 227,5903 | 93,70131 | 0,844889 | 151,284 |
| **35** | 242,1731 | 227,4395 | 93,63577 | 0,84513 | 151,0413 |
| **36** | 241,5807 | 227,6477 | 93,76563 | 0,845016 | 152,6485 |

*Tabla 7: Elección ventana fija óptima, modelo persistente probabilístico "Hoy-mañana"*

Tras observar los datos, se determina que el día que mejor compromiso obtiene entre el menor CRPS y menor RMSD, es el de 13 días atrás. Una vez se ha elegido el día, se procede a aplicar dicha ventana sobre los datos del grupo de test para obtener los resultados de la predicción probabilística.

```
# Se ha comprobado que lo mejor es 13 días atrás, lo recompongo con 13 para sacar histograma
lim_inf <- 13
cuantiles_totales <- matrix(nrow=3672, ncol=19, NA)
conta <- 1
for (ind in seq(16319, 19990, by=24 )){
  i <- ind
  gene <- Datos$ene[(ind):(ind+23)]
  ensemble <- matrix(nrow=24, ncol=36, NA)
  cuantiles <- matrix(nrow=24, ncol=19, NA)
  crps_valores <- matrix(nrow=24, ncol=1, NA)
 # Creo ensemble para el dia D. Obtengo las predicciones a base de los cuantiles con la ventana escogida
  for (j in c(1:24)){
 # Relleno la fila de la hora correspondiente del ensemble
    for (d in c(2:36)){
      k <- i+j-1-d*24
      if (k>=1) {
        ensemble[j,d] <- Datos$ene[k]
      } else {
        ensemble[j,d] <- NA
      }
    }
  }
  for (j in c(1:24)){
    cuantiles[j,] <- quantile(ensemble[j,1:lim_inf], probs = seq(0.05, 0.95, 0.05), na.rm = TRUE, names = TRUE, type = 6)
    crps_valores[j] <- crps_sample(gene[j], cuantiles[j], method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)
  }
  crps_dia[conta] <- mean(crps_valores)
  print(paste0(conta, "     ", lim_inf, "   ", crps_dia[conta]))
  prediccion[(ind-16318):(ind-16318+23)] <- cuantiles[ ,10]
  cuantiles_totales[(ind-16318):(ind-16318+23),] <- cuantiles
  conta <- conta +1
}
# Evaluo el crps, rmse, mae, mre, rix para todo el grupo de test
rank.hist <- Rankhist(cuantiles_totales, Datos$ene[16319:19990])
```

```
teo <- matrix(nrow=1, ncol=20, 3672/20)
PlotRankhist(rank.hist, mode="raw")
abline(h=3672/20, col="red",lwd=3)

file_e <- paste0("Cuantiles_persistente_13dias_mañana.xlsx")
write.xlsx(cuantiles_totales, file=file_e, sheetName = "Percentiles_test", col.names = TRUE, row.names = FALSE, append = FALSE)
```

Una vez obtenidos los percentiles de la distribución probabilista, se procede a obtener los criterios evaluadores generados por los datos del grupo de test.

```
# Programa para obtener los criterios evaluadores

options(java.parameters = "- Xmx2048m")

# -------

library(scoringRules)
library(readxl)
library(matrixStats)
library(stats)
library(KScorrect)
library(xlsx)
library(SpecsVerification)
library(quantreg)

rsq <- function (x, y) cor(x, y) ^ 2

Datos <- read_excel("Cuantiles_persistente_13dias_mañana.xlsx")

df <- data.frame(Datos)
testing <- as.matrix(df[1:3672, 1:19])
pot_test <- df[1:3672, 20]
pot_test <- as.numeric(pot_test)

crps_test <- matrix(nrow=3672, ncol=1, NA)

crps_test <- crps_sample(pot_test, testing, method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)

crps_testing <- mean(crps_test)
```

```
rmse_testing <- sqrt(mean((pot_test-testing[,10])^2))
mae_testing <- mean(abs(pot_test-testing[,10]))
rsq_testing <- rsq(pot_test,testing[,10])

rank.hist <- Rankhist(testing, pot_test)
PlotRankhist(rank.hist, mode="raw")
abline(h=3672/20, col="red",lwd=3)
teo <- matrix(nrow=1, ncol=20, 3672/20)
rmsd_testing <- sqrt(1/20*sum((rank.hist-teo)^2))

sink(paste0("Modelo_persistente_mañana_13_test.txt"))
print("\n")
print("\n")
print(paste0("CRPS test: ", crps_testing))
print(paste0("RMSE test: ", rmse_testing))
print(paste0("MAE test: ", mae_testing))
print(paste0("R2 test: ", rsq_testing))
print(paste0("RMSD test: ", rmsd_testing))
sink()
```

Los datos que se han conseguido para la ventana fija de 16 días aplicados sobre los datos del grupo de test se recogen en la siguiente tabla:

| | CRPS | RMSE | MAE | $R^2$ | RMSD |
|---|---|---|---|---|---|
| **Entrenamiento** | 215,8542 | 219,4222 | 88,6613 | 0,8522 | 202,2411 |
| **Test** | 47,76968 | 157,9422 | 61,4590 | 0,92380 | 43,51024 |

*Tabla 8: Resultados grupo de entrenamiento y test, modelo probabilístico persistente "Hoy-mañana"*

En el siguiente histograma, que corresponde a los datos del grupo de test, la línea roja determina el número de casos (en este caso sería un total de 180 observaciones aproximadamente) los cuales no deberían ser sobrepasados por ninguna de las barras siempre que se realice una predicción precisa, situación que se cumple parcialmente, proporcionando cierta homogeneidad a la muestra.

En este caso, ocurre algo similar a lo que se representaba en el histograma del modelo de “hoy-hoy” y, la predicción de la potencia producida es infravalorada en los percentiles altos y sobrevalorado en los percentiles bajos. Por tanto, ahora los valores han sufrido una ligera infravaloración, es decir, los

datos de la muestra reflejen una menor producción de la que realmente se ha predicho.

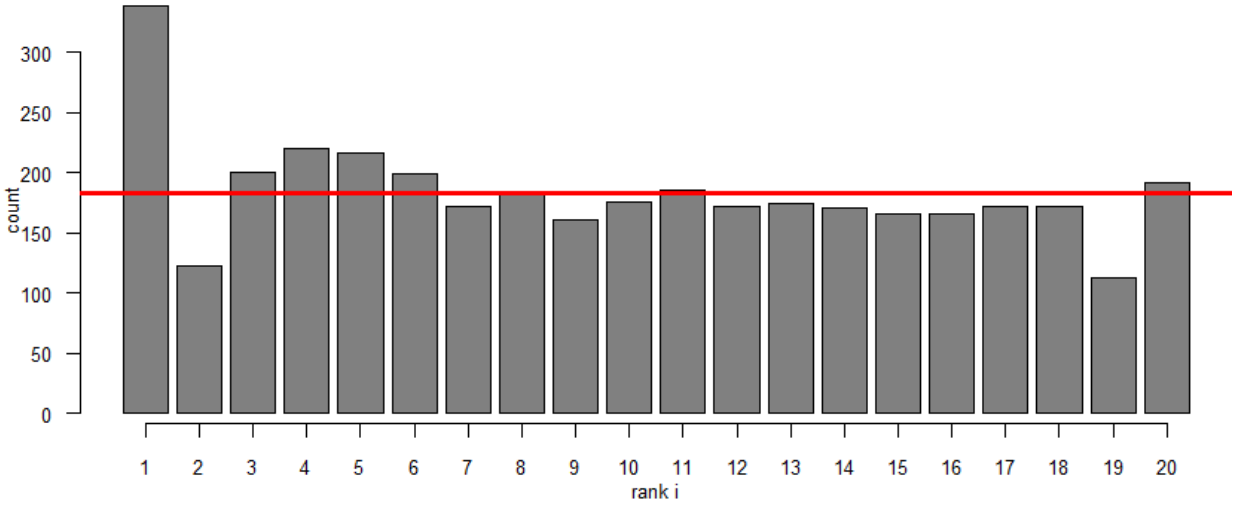


*Ilustración 15: Histograma grupo de test, modelo persistente probabilístico "Hoy-mañana"*

## 10.3 Regresión de Cuantiles

En primer lugar, explicar que es un cuantil. Generalmente es utilizado en estadística descriptiva y, se podrían definir como puntos tomados a intervalos regulares de la función de distribución de una variable aleatoria, es decir, hacen referencia a las medidas de posición no central que permiten a su vez examinar otros puntos relevantes de la distribución, los cuales no son centrales.

El término de cuantil fue utilizado por primera vez por Kendall (1940), el cual dice que el cuantil "M" de una distribución ($0 < m < 1$), sería el valor de la variable $X_m$, el cual marca un corte. De modo que, una proporción "M" de valores del grupo de muestra es igual o menor que $X_m$. Existen diferentes tipos de cuantiles, los cuales son los siguientes:

- Cuartiles: dividen la distribución en cuatro partes, que corresponden a los cuantiles 0,25; 0,50 y 0,75.
- Quintiles: seccionan la distribución en cinco partes iguales, que corresponden a los cuantiles 0,20; 0,40; 0,60 y 0,80.
- Deciles: dividen la distribución en diez partes iguales.
- Percentiles: segmentan la distribución en cien partes iguales.

Los cuantiles pueden utilizarse con diferentes tipos de variables. En el caso de realizar el cálculo con distribuciones de variable continua (por ejemplo, con datos agrupados), se conseguiría de una forma sencilla que las partes en las que se fraccione la distribución fueran todas iguales.

Sin embargo, cuando el cálculo va a realizarse con distribuciones de variable discreta (datos aislados) habrá que resignarse con que estas partes sean aproximadamente iguales. Por ello, al calcular cualquier cuantil de datos no agrupados por medio de calculadora, software o manualmente, será necesario especificar el método utilizado (debido a que hay nueve métodos diferentes para llevar a cabo el cálculo y todos ellos proporcionan resultados distintos).

La regresión de cuantiles, fue introducida por primera vez por Koenker y Basset (1982), se considera un método de estimación de la relación entre la variable endógena y los regresores alternativa a los métodos clásicos de

mínimos cuadrados ordinarios o de máxima verosimilitud. Así, mientras los procedimientos clásicos requieren unas hipótesis previas sobre la aleatoriedad de la relación, la regresión de cuantiles no necesita dichas hipótesis para la estimación de los parámetros, no considerando ninguna restricción sobre la perturbación aleatoria. El hecho de que pueda establecerse el tipo de relación entre los regresores y la endógena sin incluir ninguna hipótesis sobre la perturbación aleatoria, clasifica el método como semiparamétrico.

Además, el método de estimación mínimo cuadrático tiene por objetivo minimizar la suma de los residuos al cuadrado, mientras que en la regresión de cuantiles el objetivo es minimizar una suma de errores absolutos ponderados con pesos asimétricos.

El cuantil es un valor que minimiza una suma ponderada, donde se ponderará más la parte con menos observaciones, siendo la mediana un caso especial Q=0,50 en el que todas las observaciones tienen la misma ponderación. Habiendo definido los cuantiles incondicionales de un valor muestral, como un problema de optimización, podemos igualmente plantear los cuantiles de Y condicionados a los valores de un conjunto de regresores X.

Calcular la significatividad de los parámetros y su contraste de nulidad es más complicado en la regresión de cuantiles que en los procedimientos clásicos, ya que se trata de estimadores semiparamétricos, donde no se han establecido las hipótesis habituales sobre el término de error.

Según Buchinsky (1995), existen múltiples aplicaciones con regresiones de cuantiles y en campos muy diversos, pero en general el terreno donde dan mejores resultados frente a los procedimientos convencionales es cuando se dispone de una enorme cantidad de datos. En estos casos la información de la que se dispone no suele adaptarse a las robustas limitaciones impuestas en las hipótesis básicas del modelo de regresión lineal y sus problemas de heterocedasticidad (cuando la varianza de los errores no es constante en todas las observaciones realizadas), o asimetría son habituales. En Koenker y Hallock (2001) se detallan las ventajas de la estimación cuantílica frente a métodos más tradicionales, resaltando los casos en los que los incumplimientos de determinadas hipótesis conducen a resultados más fiables con el empleo de la regresión de cuantiles.

Asimismo, uno de los beneficios de la utilización de este tipo de estimación frente al mínimo cuadrático se produce cuando nos encontramos con elementos muestrales atípicos (outliers). Cabe resaltar, que en mínimos cuadrados todas las observaciones intervienen de igual forma y que puntos alejados o extraños del plano medio tirarán de éste, pues el objetivo es minimizar la suma de todos los residuos al cuadrado. Por el contrario, puntos atípicos en

la estimación cuantilíca, mediana, por ejemplo, no modificará la solución. Otra de las utilidades de la regresión cuantílica, es que muestra el comportamiento de los parámetros según varía el cuantil, lo que es similar a analizar la relación de las variables para diferentes valores o tamaños de la endógena estimada.

Por el contrario, una regresión de cuantiles, mostrará que en los cuantiles superiores el parámetro aumenta considerablemente de tamaño. La ventaja que aporta la regresión de cuantiles frente a los modelos clásicos, es que en cada cuantil intervienen todas las observaciones convenientemente ponderadas.

Con todo, la regresión de cuantiles se presenta como alternativa al método clásico de mínimos cuadrados ordinarios, con las ventajas e inconvenientes de no exigir el cumplimiento de las hipótesis básicas requeridas en los procedimientos clásicos, siendo menos sensible a la existencia de atípicos (presencia de outliers) en la distribución de la muestra y ofreciendo estimaciones alternativas de los parámetros cuando hay posibilidad de que aparezcan cambios de estructura en la muestra.

### 10.3.1 Ventajas de la regresión cuantílica

Algunas de las ventajas que supone la utilización de este modelo son las siguientes:

- Permite modelar los extremos de la variable respuesta.
- Permite identificar mejor el efecto de las covariables sobre la distribución condicional.
- En datos con elevada presencia de ceros, el modelado por regresión cuantílica no se ve afectada por la selección de δ en la transformación log (y + δ).
- Brinda mayor flexibilidad en el modelado de datos con altos niveles de variabilidad, describiendo el comportamiento para cada cuantil deseado.

### 10.3.2 Código para el análisis mediante la herramienta Rstudio

Código utilizado para el análisis mediante Rstudio, para la obtención de los criterios evaluadores de los modelos de predicción:

```
# Programa para realizar la regresión de cuantiles directamente con los mismos datos
# que se usaron en las predicciones determinísticas

setwd("N:/Modelos/Regresion cuantiles")

# -------

library(scoringRules)
library(readxl)
library(matrixStats)
library(stats)
library(KScorrect)
library(SpecsVerification)
library(quantreg)

rsq <- function (x, y) cor(x, y) ^ 2

Datos <- read_excel("Libro7.xlsx")

df <- data.frame(Datos)

entradas_ent <- df[1:16318,6:13]
salida_ent <- df[1:16318,3]
entradas_tst <- df[16319:19990,6:13]
salida_tst <- df[16319:19990,3]
m <- cbind(entradas_ent, salida_ent)

qtrain <- matrix(nrow=16318, ncol=19, NA)

model0.05 <- rq(salida_ent ~ .,tau=0.05, data=m)
qtrain[,1] <- predict(model0.05,newdata=entradas_ent)
model0.10 <- rq(salida_ent ~ .,tau=0.10, data=m)
qtrain[,2] <- predict(model0.10,newdata=entradas_ent)
model0.15 <- rq(salida_ent ~ .,tau=0.15, data=m)
qtrain[,3] <- predict(model0.15,newdata=entradas_ent)
```

```
model0.20 <- rq(salida_ent ~ .,tau=0.20, data=m)
qtrain[,4] <- predict(model0.20,newdata=entradas_ent)
model0.25 <- rq(salida_ent ~ .,tau=0.25, data=m)
qtrain[,5] <- predict(model0.25,newdata=entradas_ent)
model0.30 <- rq(salida_ent ~ .,tau=0.30, data=m)
qtrain[,6] <- predict(model0.30,newdata=entradas_ent)
model0.35 <- rq(salida_ent ~ .,tau=0.35, data=m)
qtrain[,7] <- predict(model0.35,newdata=entradas_ent)
model0.40 <- rq(salida_ent ~ .,tau=0.40, data=m)
qtrain[,8] <- predict(model0.40,newdata=entradas_ent)
model0.45 <- rq(salida_ent ~ .,tau=0.45, data=m)
qtrain[,9] <- predict(model0.45,newdata=entradas_ent)
model0.50 <- rq(salida_ent ~ .,tau=0.50, data=m)
qtrain[,10] <- predict(model0.50,newdata=entradas_ent)
model0.55 <- rq(salida_ent ~ .,tau=0.55, data=m)
qtrain[,11] <- predict(model0.55,newdata=entradas_ent)
model0.60 <- rq(salida_ent ~ .,tau=0.60, data=m)
qtrain[,12] <- predict(model0.60,newdata=entradas_ent)
model0.65 <- rq(salida_ent ~ .,tau=0.65, data=m)
qtrain[,13] <- predict(model0.65,newdata=entradas_ent)
model0.70 <- rq(salida_ent ~ .,tau=0.70, data=m)
qtrain[,14] <- predict(model0.70,newdata=entradas_ent)
model0.75 <- rq(salida_ent ~ .,tau=0.75, data=m)
qtrain[,15] <- predict(model0.75,newdata=entradas_ent)
model0.80 <- rq(salida_ent ~ .,tau=0.80, data=m)
qtrain[,16] <- predict(model0.80,newdata=entradas_ent)
model0.85 <- rq(salida_ent ~ .,tau=0.85, data=m)
qtrain[,17] <- predict(model0.85,newdata=entradas_ent)
model0.90 <- rq(salida_ent ~ .,tau=0.90, data=m)
qtrain[,18] <- predict(model0.90,newdata=entradas_ent)
model0.95 <- rq(salida_ent ~ .,tau=0.95, data=m)
qtrain[,19] <- predict(model0.95,newdata=entradas_ent)

qtest <- matrix(nrow=3672, ncol=19, NA)

qtest[,1] <- predict(model0.05,newdata=entradas_tst)
qtest[,2] <- predict(model0.10,newdata=entradas_tst)
qtest[,3] <- predict(model0.15,newdata=entradas_tst)
qtest[,4] <- predict(model0.20,newdata=entradas_tst)
qtest[,5] <- predict(model0.25,newdata=entradas_tst)
qtest[,6] <- predict(model0.30,newdata=entradas_tst)
qtest[,7] <- predict(model0.35,newdata=entradas_tst)
qtest[,8] <- predict(model0.40,newdata=entradas_tst)
qtest[,9] <- predict(model0.45,newdata=entradas_tst)
```

```
qtest[,10] <- predict(model0.50,newdata=entradas_tst)
qtest[,11] <- predict(model0.55,newdata=entradas_tst)
qtest[,12] <- predict(model0.60,newdata=entradas_tst)
qtest[,13] <- predict(model0.65,newdata=entradas_tst)
qtest[,14] <- predict(model0.70,newdata=entradas_tst)
qtest[,15] <- predict(model0.75,newdata=entradas_tst)
qtest[,16] <- predict(model0.80,newdata=entradas_tst)
qtest[,17] <- predict(model0.85,newdata=entradas_tst)
qtest[,18] <- predict(model0.90,newdata=entradas_tst)
qtest[,19] <- predict(model0.95,newdata=entradas_tst)

crps_horas_train <- crps_sample(salida_ent, qtrain, method = "edf", w = NULL,
bw = NULL,num_int = FALSE, show_messages = TRUE)
crps_train <- mean(crps_horas_train)
mae_train <- mean(abs(qtrain[,10]-salida_ent))
rmse_train <- sqrt(mean((qtrain[,10]-salida_ent)^2))
rsq_train <- rsq(qtrain[,10],salida_ent)
rank.hist <- Rankhist(qtrain, salida_ent)
teo <- matrix(nrow=1, ncol=20, 16318/20)
rmsd_train <- sqrt(1/20*sum((rank.hist-teo)^2))
PlotRankhist(rank.hist, mode = "raw")

crps_horas_test <- crps_sample(salida_tst, qtest, method = "edf", w = NULL, bw
= NULL,num_int = FALSE, show_messages = TRUE)
crps_test <- mean(crps_horas_test)
mae_test <- mean(abs(qtest[,10]-salida_tst))
rmse_test <- sqrt(mean((qtest[,10]-salida_tst)^2))
rsq_test <- rsq(qtest[,10],salida_tst)
rank.hist <- Rankhist(qtest, salida_tst)
teo <- matrix(nrow=1, ncol=20, 3672/20)
rmsd_test <- sqrt(1/20*sum((rank.hist-teo)^2))
PlotRankhist(rank.hist, mode = "raw")

sink(paste0("Modelo_regresion_cuantiles_meteo.txt"))
print("\n")
print("\n")
print(paste0("CRPS train: ", crps_train))
print(paste0("RMSE train: ", rmse_train))
print(paste0("MAE train: ", mae_train))
print(paste0("R2 train: ", rsq_train))
print(paste0("RMSD train: ", rmsd_train))
print(paste0("CRPS test: ", crps_test))
print(paste0("RMSE test: ", rmse_test))
print(paste0("MAE test: ", mae_test))
```

```
print(paste0("R2 test: ", rsq_test))
print(paste0("RMSD test: ", rmsd_test))
sink()
```

Los resultados obtenidos se adjuntan en la siguiente tabla, los cuales serán comparados posteriormente con el resto de modelos con la intención de determinar cuál ha realizado una mejor predicción según los criterios evaluadores. Cabe destacar que, los resultados obtenidos de la muestra de los datos de test han sido considerablemente mejores que los obtenidos del grupo de entrenamiento.

| | CRPS | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| **Entrenamiento** | 86,0312 | 222,6523 | 113,7515 | 0,8423 |
| **Test** | 79,0231 | 181,0470 | 98,7776 | 0,9189 |

*Tabla 9: Resultados grupo de entrenamiento y de test, modelo Regresión de cuantiles*

## 10.4 Regresión Lineal Múltiple

El modelo que se pretende estudiar a continuación, es uno de los más importantes y más utilizados en el campo de la estadística. Sus propiedades, utilidades y limitaciones son sobradamente conocidas. Por el computo de estos factores, añadidos a la simplicidad matemática de la relación entre la variable obtenida y las variables explicativas, son las primordiales razones de su elección.

La regresión lineal es una técnica estadística destinada a analizar las causas de por qué se producen determinados acontecimientos o como se obtienen determinados resultados. Mediante la utilización de modelos de análisis de regresión lineal múltiple se puede:

- Identificar que variables independientes (causas) explican una variable dependiente (resultado).
- Comparar y comprobar modelos causales.
- Predecir valores de una variable, es decir, a partir de unas características predecir de forma aproximada un comportamiento o estado.

Las técnicas de regresión lineal múltiple parten de (k+1) variables cuantitativas, siendo Y la variable de respuesta y ($X_1$, $X_2$, ..., $X_k$) las variables explicativas. Se basa en extender a las 'k' variables las técnicas de la regresión lineal simple. En esta línea, la variable Y se puede expresar mediante una función lineal de las variables ($X_1$, $X_2$, ..., $X_k$).

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

*Ecuación 6: Fórmula general de Regresión lineal múltiple*

Donde:

- Y: Es la variable dependiente, explicada o regresada.
- $X_k$: hace referencia a las variables explicativas, independientes o regresoras.
- $\beta_k$: son los parámetros que miden la influencia o el peso que tienen las variables explicativas sobre la variable dependiente.

Para este trabajo final de grado la ecuación que se utilizó para predecir la potencia producida tendrá la siguiente forma:

$$P_{producida_{prevista}} = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \cdots + a_6 \cdot X_6$$

*Ecuación 7:Fórmula general para calcular la potencia producida prevista, modelo de Regresión lineal múltiple*

Donde:

- $P_{producida}$: Es la variable dependiente y será la potencia prevista a producir.
- $X_k$: hace referencia a las variables explicativas, independientes o regresoras.
- $a_k$: son los parámetros que miden la influencia o el peso que tienen las variables explicativas.

Para el análisis mediante la hoja de cálculo Excel, la muestra fue fraccionada en dos partes claramente diferenciadas:

- **Grupo de entrenamiento:** serán los datos utilizados para entrenar el modelo y optimizarlo, con la finalidad de ser posteriormente evaluado por los datos correspondientes al grupo de test y validar si se obtuvo una buena predicción. Se utilizaron 16318 valores de la muestra, lo que corresponde a un total de 680 días de forma aproximada.
- **Grupo de test:** serán los datos utilizados para evaluar el modelo creado mediante el grupo de entrenamiento anteriormente descrito. Dicha muestra está formada por 3672 valores con una periodicidad de 1 hora, lo que corresponde a un total de 153 días aproximadamente.

Una vez se han elaborado ambas muestras, se procede a obtener el peso de los parámetros (los datos del grupo de entrenamiento), que permitirán determinar la influencia que tienen las diferentes variables explicativas, las cuales fueron enumeradas en apartados anteriores.

Dichos parámetros se calcularán mediante la herramienta solver. Solver es una herramienta de análisis, aplicado sobre todo en el mundo

empresarial, permitiendo calcular el valor de una celda que depende de diversos factores o variables donde a la vez existen una serie de restricciones que han de cumplirse.

Más detenidamente, lo que la herramienta Solver de Excel realiza son los cálculos para la resolución de problemas de programación lineal, en donde a partir de una función lineal a optimizar (encontrar el máximo o mínimo) y cuyas variables están sujetas a unas restricciones expresadas como inecuaciones lineales, el fin es obtener valores óptimos bien sean máximos o mínimos. En este caso, lo que se pretende es minimizar el error, con la finalidad de obtener buenos resultados predictivos, proporcionando precisión y fiabilidad.

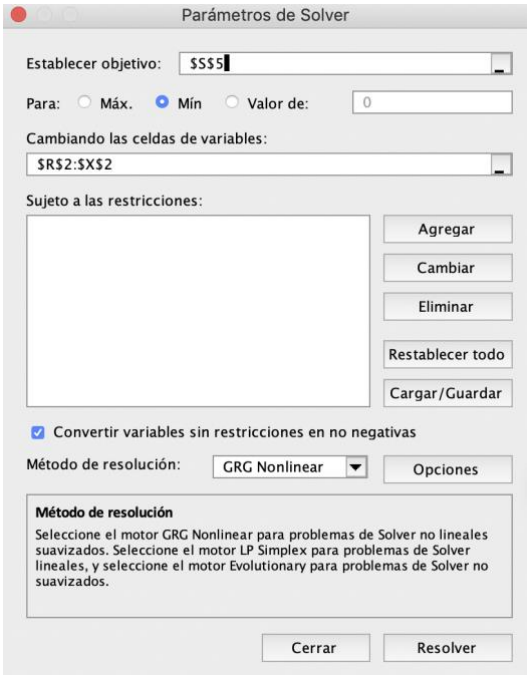


*Ilustración 16: Ejemplo de la resolución mediante la herramienta Solver*

Después de aplicar la herramienta Solver, los resultados obtenidos sobre los distintos parámetros son los siguientes:

| $a_0$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ |
|---|---|---|---|---|---|---|
| 11,4324761 | 0 | 1,70580839 | 5,25955853 | 0 | 53,9477679 | 0 |

*Tabla 10: Parámetros para determinar la ecuación de cálculo de la potencia producida prevista*

Se observa que hay parámetros que no tiene ninguna importancia ($a_1$, $a_4$ y $a_6$) y lo parámetros con más pesos de orden creciente a decreciente son los siguientes:

1. **$a_5$:** Fracción total de nubes
2. **$a_0$:** Temperatura
3. **$a_3$:** Velocidad del viento
4. **$a_2$:** Radiación media

Una vez obtenidos los parámetros, se obtiene la siguiente ecuación, mediante la cual se obtendrá un primer modelo y se podrá observar como se ajusta a los datos de entrenamiento.

$$P_{producida_{prevista}} = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \cdots + a_6 \cdot X_6$$

*Ecuación 8:Fórmula general para calcular la potencia prevista, modelo de Regresión lineal múltiple*

$$P_{producida_{prevista}} = 11{,}4325 + 1{,}7058 \cdot X_2 + 5{,}2596 \cdot X_3 + 53{,}9477 \cdot X_5$$

*Ecuación 9:Fórmula parametrizada con los valores obtenidos de Solver, modelo de Regresión lineal múltiple*

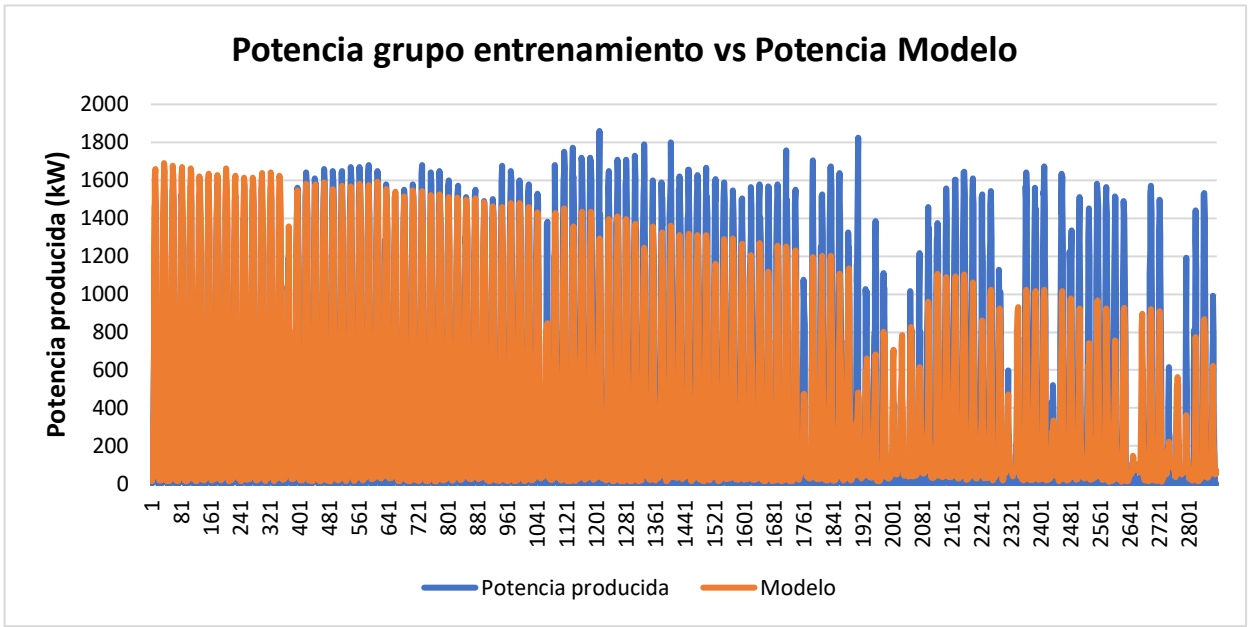


*Ilustración 17:Representación de la potencia del grupo de entrenamiento frente a la potencia del modelo, modelo Regresión lineal múltiple*

Puede observarse, que el modelo de predicción no se ajusta completamente al grupo de entrenamiento.

A continuación, habiendo elaborado el modelo con los datos correspondientes al grupo de entrenamiento, se procederá a ser aplicado sobre los datos del grupo de test, los cuales permitirán evaluar la precisión predictiva del modelo.

Se procede a ejecutar el cálculo de los percentiles para poder estudiar como han sido los resultados obtenidos (la información acerca de los percentiles se encuentra añadida en el apartado del modelo climatológico).

Una vez se han realizados los pasos previos correspondientes, se procede a realizar el cálculo de los mismos (desde 0,05 hasta 0,95 en intervalos de 0,05). Para ello, se utiliza una función de Excel, INV.NORM (devuelve el inverso de la distribución normal acumulativa para la media y desviación estándar especificadas), la cual utiliza la desviación estándar obtenida anteriormente con los datos del grupo de entrenamiento y los datos logrados, después de evaluar el modelo de predicción con los datos del grupo de test.

En la siguiente tabla, se muestra un ejemplo de los percentiles calculados.

| 0,95 | 0,75 | 0,5 | 0,25 | 0,05 |
|---|---|---|---|---|
| 381,345584 | 169,30276 | 21,9140171 | 0 | 0 |
| 378,303917 | 166,261093 | 18,8723506 | 0 | 0 |
| 383,218505 | 171,175681 | 23,7869381 | 0 | 0 |
| 383,284652 | 171,241828 | 23,8530852 | 0 | 0 |
| 382,438389 | 170,395565 | 23,0068225 | 0 | 0 |
| 379,367274 | 167,32445 | 19,9357074 | 0 | 0 |
| 375,457147 | 163,414323 | 16,0255803 | 0 | 0 |
| 374,157669 | 162,114845 | 14,7261021 | 0 | 0 |
| 415,273843 | 203,231019 | 55,8422769 | 0 | 0 |
| 563,184605 | 351,14178 | 203,753038 | 56,3642956 | 0 |
| 894,796669 | 682,753845 | 535,365103 | 387,97636 | 175,933536 |
| 1328,82924 | 1116,78641 | 969,397671 | 822,008929 | 609,966104 |
| 1623,75105 | 1411,70822 | 1264,31948 | 1116,93074 | 904,887913 |
| 1849,24648 | 1637,20366 | 1489,81492 | 1342,42617 | 1130,38335 |
| 1842,58936 | 1630,54654 | 1483,1578 | 1335,76905 | 1123,72623 |
| 1710,9566 | 1498,91317 | 1351,52443 | 1204,13569 | 992,092865 |
| 1727,67045 | 1515,62762 | 1368,23888 | 1220,85014 | 1008,80731 |
| 1643,20431 | 1431,16149 | 1283,77275 | 1136,384 | 924,341181 |
| 1501,17376 | 1289,13094 | 1141,7422 | 994,353455 | 782,310631 |
| 1232,49051 | 1020,44769 | 873,058944 | 725,670202 | 513,627377 |
| 902,487504 | 690,44468 | 543,055937 | 395,667195 | 183,624371 |

| | | | | |
|---|---|---|---|---|
| 603,994731 | 391,951907 | 244,563165 | 97,1744222 | 0 |
| 395,554371 | 183,511547 | 36,1228048 | 0 | 0 |
| 388,326935 | 176,284111 | 28,8953685 | 0 | 0 |
| 388,643028 | 176,600204 | 29,2114612 | 0 | 0 |
| 391,88391 | 179,841086 | 32,4523436 | 0 | 0 |
| 404,183184 | 192,14036 | 44,7516173 | 0 | 0 |
| 408,852277 | 196,809453 | 49,4207104 | 0 | 0 |
| 402,519549 | 190,476725 | 43,0879827 | 0 | 0 |
| 393,968311 | 181,925486 | 34,536744 | 0 | 0 |
| 385,752147 | 173,709323 | 26,3205802 | 0 | 0 |
| 384,565321 | 172,522497 | 25,1337543 | 0 | 0 |
| 389,957213 | 177,914389 | 30,5256461 | 0 | 0 |
| 487,009786 | 274,966962 | 127,578219 | 0 | 0 |
| 609,492859 | 397,450034 | 250,061292 | 102,67255 | 0 |
| 852,931608 | 640,888784 | 493,500042 | 346,111299 | 134,068475 |
| 985,95567 | 773,912846 | 626,524103 | 479,135361 | 267,092537 |
| 1006,98232 | 794,939497 | 647,550754 | 500,162012 | 288,119187 |
| 1274,0545 | 1062,01168 | 914,622933 | 767,23419 | 555,191366 |
| 1500,53878 | 1288,49595 | 1141,10721 | 993,71847 | 781,675646 |
| 1462,32274 | 1250,27991 | 1102,89117 | 955,502429 | 743,459605 |
| 1396,2029 | 1184,16008 | 1036,77134 | 889,382594 | 677,33977 |
| 1053,53414 | 841,491317 | 694,102574 | 546,713832 | 334,671008 |
| 670,502091 | 458,459267 | 311,070525 | 163,681782 | 0 |
| 577,009183 | 364,966359 | 217,577617 | 70,1888741 | 0 |
| 440,011127 | 227,968302 | 80,57956 | 0 | 0 |
| 420,012239 | 207,969415 | 60,5806721 | 0 | 0 |
| 412,230945 | 200,188121 | 52,7993784 | 0 | 0 |

***Tabla 11: Ejemplo de percentiles calculados mediante hoja de cálculo Excel, modelo Regresión lineal múltiple***

A continuación, se representan tres días donde se puede apreciar como se ajusta la potencia real producida frente a los percentiles calculados derivados del modelo de predicción probabilístico.

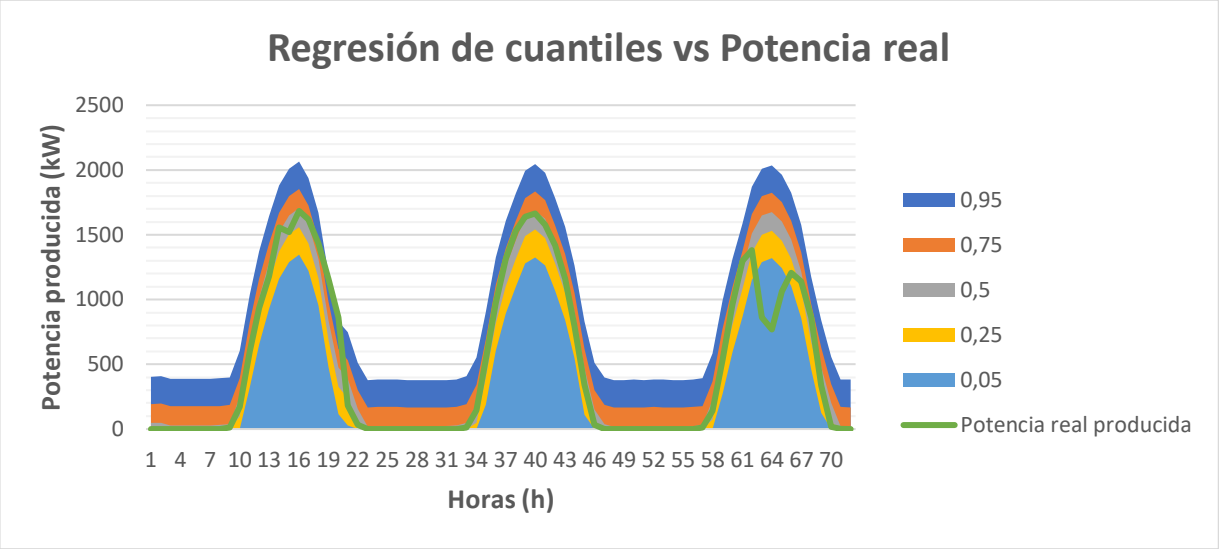


*Ilustración 18: Comparativa entre los percentiles obtenidos frente a la potencia real producida, modelo Regresión lineal múltiple*

Se aprecia que la potencia real producida se ajusta bastante bien al modelo de predicción creado, puesto que, adquiere el valor del percentil 50 (la mediana), el cual nos proporciona la información de que el 50 % de los valores se sitúa por debajo de esa estimación; de un valor aproximado de 1700 kW. Dado que el valor pico que tiene la planta fotovoltaica es de 2160 kW se podría decir que se ajusta adecuadamente proporcionando información fiable y precisa.

En el primer gráfico se puede apreciar la como se ajustan los valores predichos por el modelo a una línea de tendencia, observándose que el núcleo principal de valores se sitúa entorno a esa línea de regresión lineal. Adjuntándose la ecuación de la recta y el coeficiente de determinación (indicando este último un alto grado de precisión en los datos resultantes).

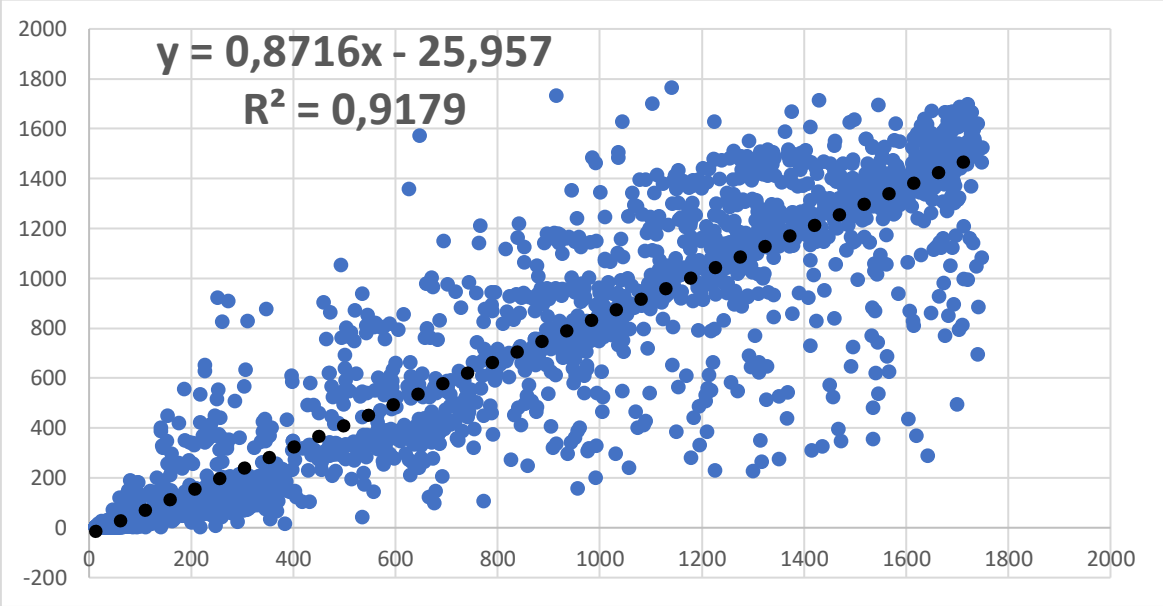


*Ilustración 19: Ajuste línea de regresión entre los valores reales y los predichos*

En el segundo gráfico que se mostrará a continuación, se puede analizar con claridad como cambia el acoplamiento del modelo de predicción probabilístico. Se remarca este hecho puesto que, en la gráfica anterior en la que se mostraba como se ajustaba el grupo de entrenamiento al modelo, no había tanta "afinidad" como la que puede apreciarse entre el grupo de test y el modelo de predicción. Por tanto, se concluye que el modelo elaborado predice con una fiabilidad y precisión elevada.

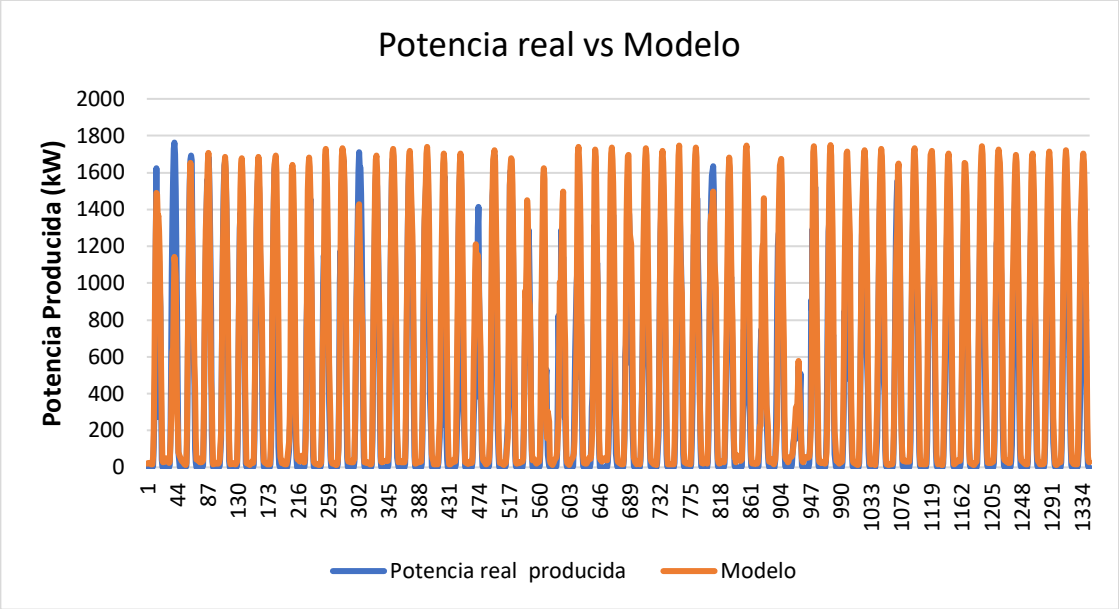


*Ilustración 20: Comparativa de la potencia real producida frente a la previsión del modelo de Regresión lineal múltiple*

### 10.4.1 Código para el análisis mediante la herramienta Rstudio

Código utilizado para el análisis mediante Rstudio, para la obtención de los criterios evaluadores de los modelos de predicción:

```
# Programa para calcular los criterios evaluadores pertinentes

#options(java.parameters = "- Xmx2048m")

setwd("N:/Modelos/Regresion lineal múltiple/crps")

# -------

library(scoringRules)
```

```
library(readxl)

rsq <- function (x, y) cor(x, y) ^ 2

Datos <- read_excel("Libro5.xlsx")

cuantiles <- as.matrix(Datos[1:3672,4:22])
real <- as.numeric(as.matrix(Datos[1:3672,2]))
prevista <- as.matrix(Datos[1:3672,1])

crps_valores <- crps_sample(real, cuantiles, method = "edf", w = NULL, bw =
NULL,num_int = FALSE, show_messages = TRUE)
rmse <- sqrt(mean((real-prevista)^2))
mae <- mean(abs(real-prevista))
rsq_testing <- rsq(prevista,real)

valor_crps <- mean(crps_valores)
```

Los resultados obtenidos se adjuntan en la siguiente tabla, los cuales serán comparados posteriormente con el resto de modelos con la intención de determinar cuál ha realizado una mejor predicción según los criterios evaluadores.

| | **CRPS** | **RMSE** | **MAE** | **$R^2$** |
|---|---|---|---|---|
| ***Entrenamiento*** | 102,1798 | 218,5122 | 130,5623 | 0,8437 |
| ***Test*** | 93,8734 | 200,4418 | 125,3817 | 0,9179 |

*Tabla 12: Resultados grupo de entrenamiento y de test, modelo Regresión lineal múltiple*

## 10.5 Random Forest (Bosques aleatorios)

### 10.5.1 Árboles de clasificación

En primer lugar, para poder explicar el funcionamiento o la forma que tienen los bosques aleatorios de analizar las muestras, se procederá a explicar de forma somera que son los árboles de clasificación.

Los árboles de clasificación proporcionan un enfoque de clasificación supervisada, el hecho de utilizar un árbol surge de su estructura, es decir, un árbol se compone de una raíz, nodos (las posiciones donde las ramas sufren bifurcaciones), ramas y hojas; asimismo, un árbol de clasificación se construye de forma similar, a partir de nodos que representan los círculos y las ramas son representadas por los segmentos que conectan los nodos.

Un árbol de clasificación se inicia desde la raíz, se extiende hacia abajo y normalmente se diseña desde la izquierda hasta la derecha. El nodo inicial se llama nodo raíz, mientras los nodos en los extremos de la cadena se les conocen como nodos hoja.

### 10.5.2 Bagging y boosting

Bagging y Boosting se enfocan en usar modelos individuales para desarrollar un mejor modelo predictivo. En Bagging se da a los modelos un peso igual a todos los atributos, mientras que el Boosting; se dan distintos pesos a los modelos con el objetivo de proporcionar mayor ponderación a aquellos que resaltan más.

Lo que se podría pensar en un principio, es que los árboles que pertenecen a una misma muestra fueran casi idénticos, por lo tanto, se obtendría una misma predicción para una nueva situación; pero esta suposición no es cierta, más aún si los conjuntos de datos son reducidos.

Si los datos del grupo de entrenamiento sufrieran algún cambio, entonces se tendría como resultado fácilmente un atributo diferente, lo cual implica que existe la posibilidad de que en los casos de prueba para algunos árboles de decisión se produzcan predicciones correctas y otras no.

Ambos son métodos de aprendizaje estadístico cuyo procedimiento tiene como propósito la reducción de la varianza.

### 10.5.3 Concepto Random Forest

Una vez comprendido el funcionamiento de los árboles de clasificación, se puede explicar de forma genérica las bases que constituyen el Random Forest. El Random Forest es una combinación de árboles predictivos, es decir, una modificación del Bagging (da a los modelos un peso igual), el cual trabaja con una colección de árboles relativamente no correlacionados y realiza el promedio de los resultados obtenidos, en el cual se tiene que cada árbol depende de los valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque.

Previamente, se crea un conjunto de árboles, de forma que la nueva predicción se aplica a cada uno de los árboles creados anteriormente, y mediante un proceso de "votación mayoritaria", a la predicción le es asignada la clase que más votos haya recibido. Cabe destacar, que los árboles no se construyen todos de la misma forma, por lo que pueden proporcionar resultados distintos.

Ahondando en la forma en la que el bosque es construido, cada árbol de clasificación se elabora de la siguiente manera:

- Con un tamaño de muestra X, se escogen de esa misma muestra X datos de forma aleatoria, pero realizando reemplazamiento. Dicho procedimiento, seguido de selección de muestras y aplicando un reemplazamiento es lo que se denomina como Bootstrapping, un método muy utilizado en bosques aleatorios.

- Dadas Z variables de entrada, se elige un número inferior z << Z de variables para que en cada nodo se vayan seleccionando nuevas de forma aleatoria para examinar la mejor partición de cada nodo. El número de variables z será constante el todo el proceso que lleve a cabo el bosque aleatorio.

- Se genera cada árbol sin podar hasta la máxima extensión que sea posible.

Teniendo en cuenta lo enumerado anteriormente, la tasa de error de un Random Forest es dependiente de dos parámetros:

- La correlación existente entre los árboles. La forma que tiene este parámetro de influir en el error de predicción del Random Forest es la siguiente, a mayor correlación entre los distintos árboles del

bosque, mayor será el error que obtenga el algoritmo. Dicha discrepancia se soluciona a través de la aleatorización con la que se escogen las diferentes variables explicativas.

- La importancia de cada árbol de forma individual, esto se traduce en que la existencia de árboles que proporcionen mucha información beneficia a la capacidad predictiva del Random Forest.

Hay que tener en cuenta que, la influencia de estos dos parámetros está estrechamente ligada con el número de variables de entrada, es decir, si se reduce el número de variables de entrada mejorará a la no correlación entre los árboles, pero a su vez será menos probable que se aporte mayor información de manera general.

Algunas de las razones por las que se utiliza tan ampliamente el Random Forest son las siguientes:

- Puede manejar enormes cantidades de datos.
- Puede arrojar resultados de importancia de variables en la clasificación.
- Es un algoritmo muy preciso.

**Ejemplo simplificado del funcionamiento del bosque aleatorio**

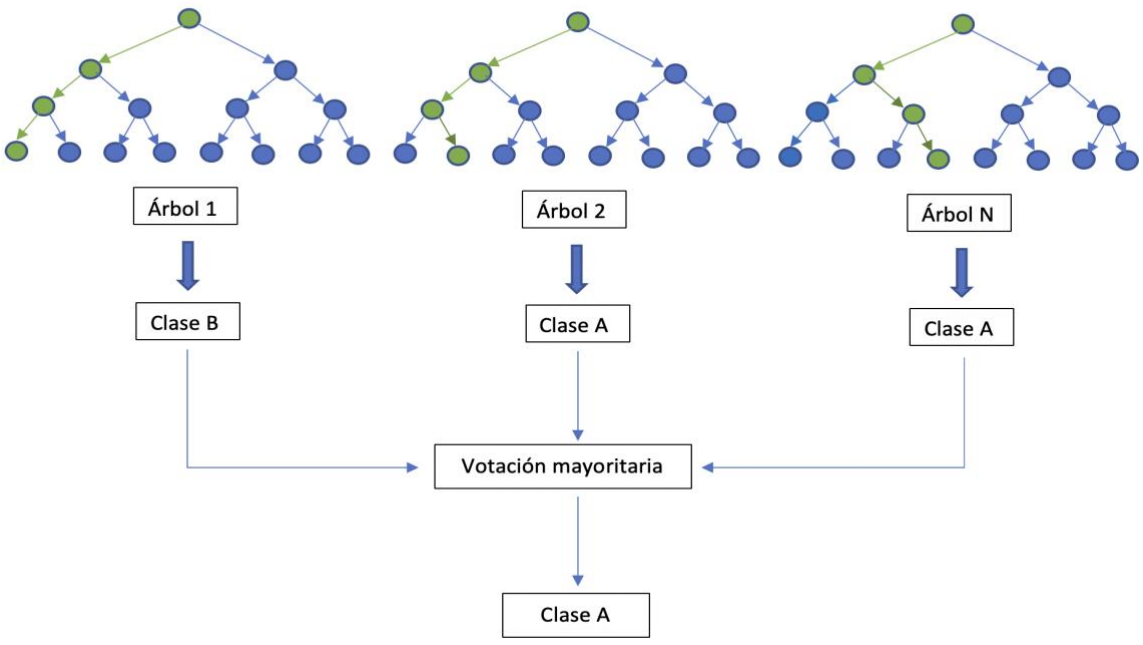


*Ilustración 21: Ejemplo del funcionamiento del Random Forest*

Este modelo de predicción fue realizado mediante el software libre R. A continuación, se adjuntará el programa utilizado para llevar a cabo el modelo de predicción. Se ha realizado con 2 - 6 variables y 3 - 50 nodos, permitiendo así un gran abanico de combinaciones posibles.

```
# Programa para realizar Random Forest directamente con los mismos datos
# que se usaron en las predicciones determinísticas

setwd("N:/Modelos/quantregForest")

# -------

library(scoringRules)
library(readxl)
library(matrixStats)
library(stats)
library(KScorrect)
library(SpecsVerification)
library(quantreg)
library(doParallel)
library(quantregForest)
library(xlsx)

rsq <- function (x, y) cor(x, y) ^ 2
Datos <- read_excel("Datos_regresion_lineal.xlsx")

df <- data.frame(Datos)

entradas_ent <- df[1:16318,6:13]
salida_ent <- df[1:16318,3]
entradas_tst <- df[16319:19990,6:13]
salida_tst <- df[16319:19990,3]

nodes <- 4 # probar con 3 a 50
mm <- 3 # probrar entre 2 y 6

qrf <- quantregForest(x=entradas_ent, y=salida_ent, nthreads=4,
nodesize=nodes, mtry=mm)conditionalQuantiles  <- predict(qrf,  entradas_ent,
what = seq(0.05, 0.95, 0.05))

# file_e <- paste0("Resultados.xlsx")
# write.xlsx(conditionalQuantiles, file=file_e, sheetName = "Percentiles_test",
col.names = TRUE, row.names = FALSE, append = FALSE)
```

```
# conditionalQuantiles  <- predict(qrf,  entradas_ent, what = seq(0.05, 0.95, 0.05))
# write.xlsx(conditionalQuantiles, file=file_e, sheetName = "Percentiles_train", col.names = TRUE, row.names = FALSE, append = TRUE)


#Calculo el valor de los errores de predicción para la muestra utilizada en el entrenamiento
crps_valores <- crps_sample(salida_ent, conditionalQuantiles, method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)
crps_train <- mean(crps_valores)
prevista <- conditionalQuantiles[,10]
rmse_train <- sqrt(mean((salida_ent-prevista)^2))
mae_train <- mean(abs(salida_ent-prevista))
rsq_train <- rsq(prevista,salida_ent)


#Calculo el valor de los errores de predicción para la muestra utilizada en el test
conditionalQuantiles  <- predict(qrf,  entradas_tst, what = seq(0.05, 0.95, 0.05))
crps_valores <- crps_sample(salida_tst, conditionalQuantiles, method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)
crps_test <- mean(crps_valores)
prevista <- conditionalQuantiles[,10]
rmse_test <- sqrt(mean((salida_tst-prevista)^2))
mae_test <- mean(abs(salida_tst-prevista))
rsq_test <- rsq(prevista,salida_tst)

rank.hist <- Rankhist(conditionalQuantiles, salida_tst)
teo <- matrix(nrow=1, ncol=20, 3672/20)
PlotRankhist(rank.hist, mode="raw")
abline(h=3672/20, col="red",lwd=3)
rmsd_test <- sqrt(1/20*sum((rank.hist-teo)^2))
```

En la siguiente tabla se muestran los resultados obtenidos tras la ejecución del programa, dicha tabla contiene los indicadores estadísticos elegidos para determinar la fiabilidad, precisión y nitidez de la predicción probabilística.

| Nº nodos | Nº variables | CRPS entren | RMSE entren | MAE entren | RSQ entren | CRPS test | RMSE test | MAE test | RSQ test | RMSD test |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 9,3284 | 7,2328 | 0,5214 | 0,9998 | 52,3234 | 150,3100 | 70,8525 | 0,9356 | 54,7589 |
| 3 | 3 | 8,7221 | 7,9841 | 0,4516 | 0,9998 | 52,6020 | 151,1488 | 70,8114 | 0,9349 | 52,0014 |
| 3 | 4 | 8,6094 | 9,9214 | 0,5628 | 0,9997 | 52,8864 | 152,4815 | 71,3374 | 0,9342 | 55,1411 |
| 3 | 5 | 8,5367 | 7,4602 | 0,4407 | 0,9998 | 53,2655 | 152,9522 | 71,6839 | 0,9338 | 57,5746 |
| 3 | 6 | 8,5197 | 12,6872 | 0,6111 | 0,9994 | 53,4750 | 153,0518 | 71,6341 | 0,9337 | 54,8109 |
| 4 | 2 | 11,7101 | 25,6122 | 3,7682 | 0,9978 | 52,4283 | 150,4775 | 70,837 | 0,9353 | 54,3354 |
| 4 | 3 | 10,8676 | 21,9193 | 2,9324 | 0,9984 | 52,6431 | 151,5386 | 71,0676 | 0,9348 | 54,4264 |
| 4 | 4 | 10,6571 | 23,6645 | 3,0326 | 0,9981 | 52,8066 | 151,8903 | 71,2553 | 0,9346 | 56,7876 |
| 4 | 5 | 10,4858 | 24,2375 | 3,0016 | 0,9981 | 53,1404 | 152,6293 | 71,5391 | 0,9342 | 52,9201 |
| 4 | 6 | 10,4151 | 22,2789 | 2,8183 | 0,9984 | 53,5099 | 153,7174 | 72,1629 | 0,9334 | 54,5907 |
| 5 | 2 | 13,9464 | 38,2338 | 7,9741 | 0,9952 | 52,2451 | 150,305 | 70,7034 | 0,9353 | 52,2507 |
| 5 | 3 | 12,9527 | 37,4936 | 7,0866 | 0,9954 | 52,6465 | 151,4201 | 70,8322 | 0,9349 | 56,5299 |
| 5 | 4 | 12,5905 | 35,8089 | 6,6681 | 0,9958 | 52,7336 | 151,7736 | 71,1971 | 0,9348 | 53,8009 |
| 5 | 5 | 12,3891 | 34,2607 | 6,3915 | 0,9961 | 53,0404 | 152,5518 | 71,4437 | 0,9341 | 54,8948 |
| 5 | 6 | 12,2754 | 34,9254 | 6,4391 | 0,9959 | 53,2984 | 152,5165 | 71,5938 | 0,9343 | 57,4459 |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| 48 | 2 | 43,8429 | 135,2472 | 60,1511 | 0,9398 | 53,7576 | 150,7747 | 72,5966 | 0,9358 | 56,0896 |
| 48 | 3 | 42,2451 | 132,6675 | 58,8021 | 0,9421 | 53,2578 | 150,9933 | 71,7964 | 0,9354 | 55,6986 |
| 48 | 4 | 41,6444 | 132,4539 | 58,6016 | 0,9422 | 52,9755 | 151,3557 | 71,6782 | 0,9348 | 53,4194 |
| 48 | 5 | 41,2732 | 131,8149 | 58,0862 | 0,9428 | 53,1027 | 152,0421 | 71,7641 | 0,9346 | 56,5839 |
| 48 | 6 | 41,0911 | 131,7553 | 58,1735 | 0,9429 | 53,5957 | 152,8484 | 72,6326 | 0,9346 | 54,8484 |
| 49 | 2 | 44,2563 | 135,9065 | 60,9048 | 0,9392 | 53,4349 | 150,0713 | 72,1482 | 0,9359 | 58,8000 |
| 49 | 3 | 42,4724 | 133,2361 | 59,0933 | 0,9416 | 53,2572 | 151,2403 | 71,8134 | 0,9352 | 55,9655 |
| 49 | 4 | 41,7282 | 132,1523 | 58,5841 | 0,9425 | 52,9963 | 150,9322 | 71,4168 | 0,9352 | 49,8979 |
| 49 | 5 | 41,4338 | 132,0648 | 58,3868 | 0,9426 | 53,3033 | 152,1652 | 71,9608 | 0,9347 | 50,3243 |
| 49 | 6 | 31,3243 | 132,5181 | 58,5866 | 0,9422 | 53,4518 | 152,5048 | 72,3241 | 0,9344 | 55,5782 |
| 50 | 2 | 44,3851 | 136,0438 | 60,8909 | 0,9391 | 53,3799 | 149,4731 | 71,7896 | 0,9362 | 55,4899 |
| 50 | 3 | 42,7280 | 133,8635 | 59,4239 | 0,9411 | 53,1858 | 150,7595 | 71,8131 | 0,9354 | 57,3527 |
| 50 | 4 | 42,0911 | 133,5178 | 59,1370 | 0,9413 | 53,1028 | 151,3671 | 71,6262 | 0,9351 | 54,9193 |
| 50 | 5 | 41,5861 | 132,4042 | 58,5186 | 0,9423 | 53,2658 | 152,1969 | 71,8497 | 0,9347 | 52,5836 |
| 50 | 6 | 41,5045 | 132,8434 | 58,7298 | 0,9419 | 53,1986 | 152,5962 | 71,9866 | 0,9340 | 51,5621 |

*Tabla 13: Ejemplo de los datos obtenidos mediante Rstudio en el primer intento, modelo Random Forest*

Se ha realizado una representación gráfica del CRPS, MAE y $R^2$, de los datos del grupo de entrenamiento frente a los datos del grupo de test, para observar la tendencia de ambos y ver como llegaban a un punto en el que no podían ajustarse más.

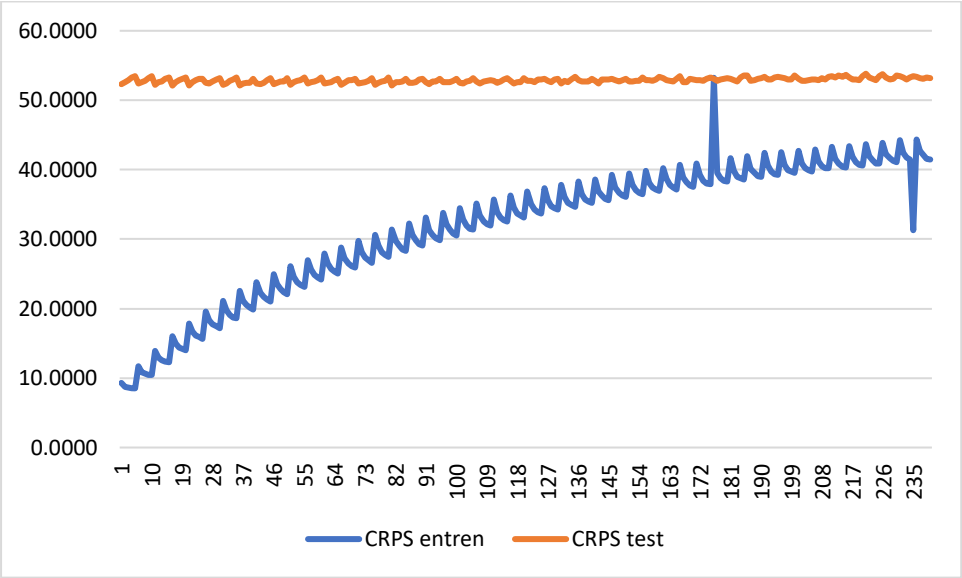


*Ilustración 22: Representación del CRPS del grupo de entrenamiento frente al CRPS del grupo de test, modelo Random Forest*

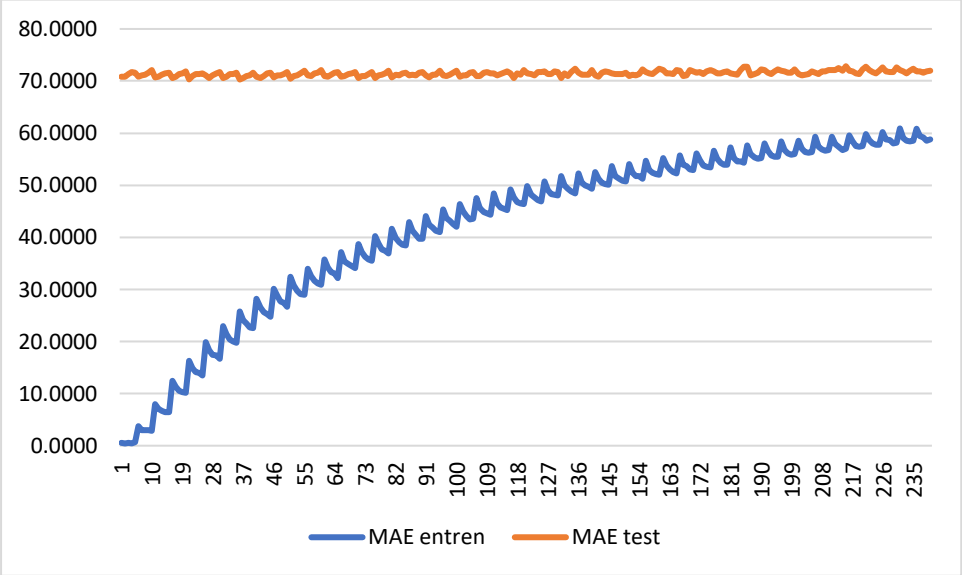


*Ilustración 23: Representación del MAE del grupo de entrenamiento frente al MAE del grupo de test, modelo Random Forest*

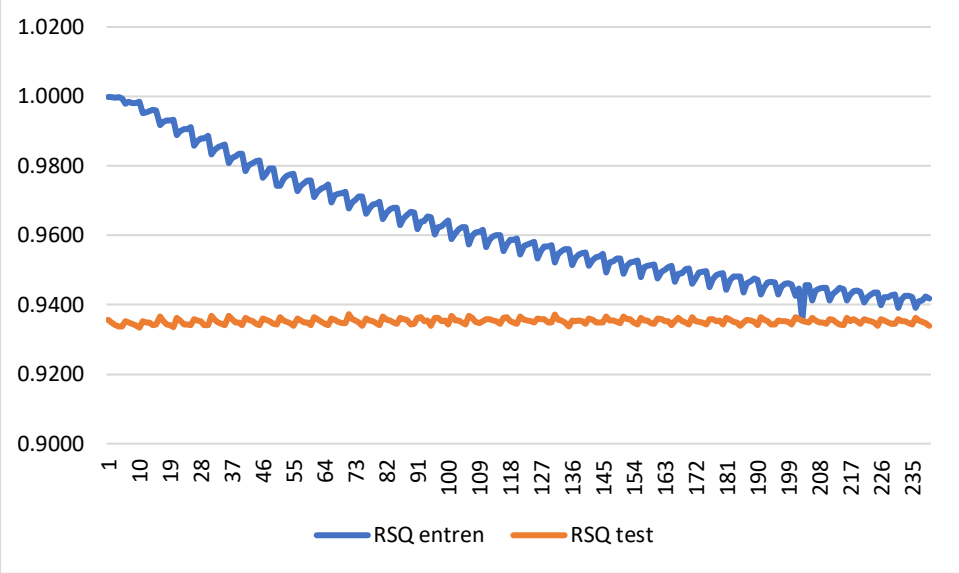


*Ilustración 24: Representación del $R^2$ del grupo de entrenamiento frente al $R^2$ del grupo de test, modelo Random Forest*

Después de analizar los datos obtenidos del algoritmo de predicción, se detecta que se ha producido un *overfitting* (sobreajuste), término que se utiliza generalmente en algoritmos de *Machine Learning* y estadística.

Sin embargo, dependiendo del tipo de modelo que se ajuste a un grupo de variables, es probable obtener un modelo que se ajuste con mucha precisión a los datos utilizados en el grupo de entrenamiento, los cuales serían los datos de partida, pero que no pueda predecir con tanta precisión y fiabilidad para otro grupo de datos, utilizados en este caso para evaluar la calidad de la predicción generada por el algoritmo.

Existen diversas técnicas que permiten subsanar este problema común, como es la poda de árboles, la validación cruzada… Cabe destacar, en el Random Forest, por la forma en la que se encuentra diseñado el algoritmo, es improbable que se produzca el fenómeno de overfitting, aunque, dependiendo del número de variables que se empleen para la construcción de los árboles o del número de árboles que se creen, existe la posibilidad de que se produzca este fenómeno. Por tanto, para subsanar el overfitting que se ha originado sobre los datos del grupo de entrenamiento se va a utilizar la validación cruzada.

### 10.5.4 Validación cruzada (Cross validation)

La validación cruzada es una técnica utilizada cuando se pretende manejar una cantidad elevada de valores, cuya finalidad será garantizar la independencia de los resultados que se obtenidos a la hora de realizar una partición de los datos empleados como grupo de entrenamiento y grupo de test.

Dicha técnica se basa en una variación del *Holdout method*, donde los datos que se utilizarán son divididos en dos grupos claramente diferenciados, los grupos mencionados anteriormente de entrenamiento y test. Por tanto, el modelo que se pretende elaborar se realizará exclusivamente con los datos que constituyen el grupo de entrenamiento, con la finalidad de verificar los resultados obtenidos de las predicciones con los datos del grupo de test.

El *Holdout method* es una técnica práctica a la hora de analizar datos, pero resulta insuficiente por sí misma, debido a que no se puede garantizar la independencia a la hora de elegir las variables con las que se elaborará el modelo. Para subsanar el problema, se realiza una aleatorización de la muestra antes de elegir qué conjuntos de variables explicativas formarán parte del grupo del test de entrenamiento.

Para asegurar la autodeterminación de los resultados obtenidos de los datos de entrada, se utilizará la validación cruzada. Existen dos tipos de validación cruzada, siendo el segundo de ellos el que más se utiliza, debido a su sencillez y claridad.

Los dos tipos son los siguientes:

- ***Validación cruzada aleatoria:*** este procedimiento difiere con el comentado anteriormente en que ahora, en vez de aleatorizar los datos y realizar una división en secciones, lo que se hace en primer lugar es aleatorizar y posteriormente, se escogerán los datos que vayan a formar parte del grupo de entrenamiento, que será también de forma aleatoria. Asimismo, será preciso definir el número de datos que formarán parte de cada conjunto o grupo de entrenamiento y determinar a su vez el número de iteraciones que se realizarán. Surge un problema al utilizar esta metodología, y es la existencia de una alta probabilidad de superponer valores del grupo de entrenamiento, y será muy posible que haya muestras que se hayan evaluado diversas veces y, por el contrario, otras que no hayan sido utilizadas en ninguna de las iteraciones realizadas.

- ***K-fold cross validation:*** dicho método consiste en realizar una división de la muestra completa en K subgrupos, de forma que en cada ocasión se elegirá a uno de esos subgrupos como grupo de test y utilizando el resto de subgrupos como grupo de entrenamiento para poder realizar la predicción. El proceso se deberá repetir K veces, cambiando consecuentemente el subgrupo que constituirá el grupo de test. Será necesario generar K veces más modelos de predicción por lo que conllevará un esfuerzo mayor.

En el presente trabajo final de grado se ha decido utilizar la K-fold cross validation, debido a que es más sencillo y evita alguno de los problemas existentes en el otro método.

Para el análisis mediante Rstudio, la muestra fue fraccionada en dos partes claramente diferenciadas:

- **Grupo de entrenamiento:** serán los datos utilizados para entrenar el modelo y optimizarlo, con la finalidad de ser posteriormente evaluado por los datos correspondientes al grupo de test y validar si se obtuvo una buena predicción. Se utilizaron 16318 valores de la muestra, lo que corresponde a un total de 680 días de forma aproximada. Los datos que constituyen este grupo fueron ordenados de forma aleatoria.

- **Grupo de test:** serán los datos utilizados para evaluar el modelo creado mediante el grupo de entrenamiento anteriormente descrito. Dicha muestra está formada por 3672 valores, lo que corresponde a un total de 153 días aproximadamente.

Una vez se han elaborado ambas muestras, internamente, en el grupo de entrenamiento se realizará una división en subgrupos, aplicando el método de *K-fold cross validation*, actuando de la siguiente forma:

1. Se realizará la división en subgrupos del grupo de entrenamiento en 3 partes.
2. Una vez elaborados los subgrupos, dependiendo de la combinación que se realice, se obtendrán tres grupos diferentes, donde, dos de los subgrupos creados anteriormente se utilizarán como grupo de entrenamiento y el restante se utilizará para evaluar el modelo de predicción. Dicha explicación se interpretará mejor en la siguiente imagen.

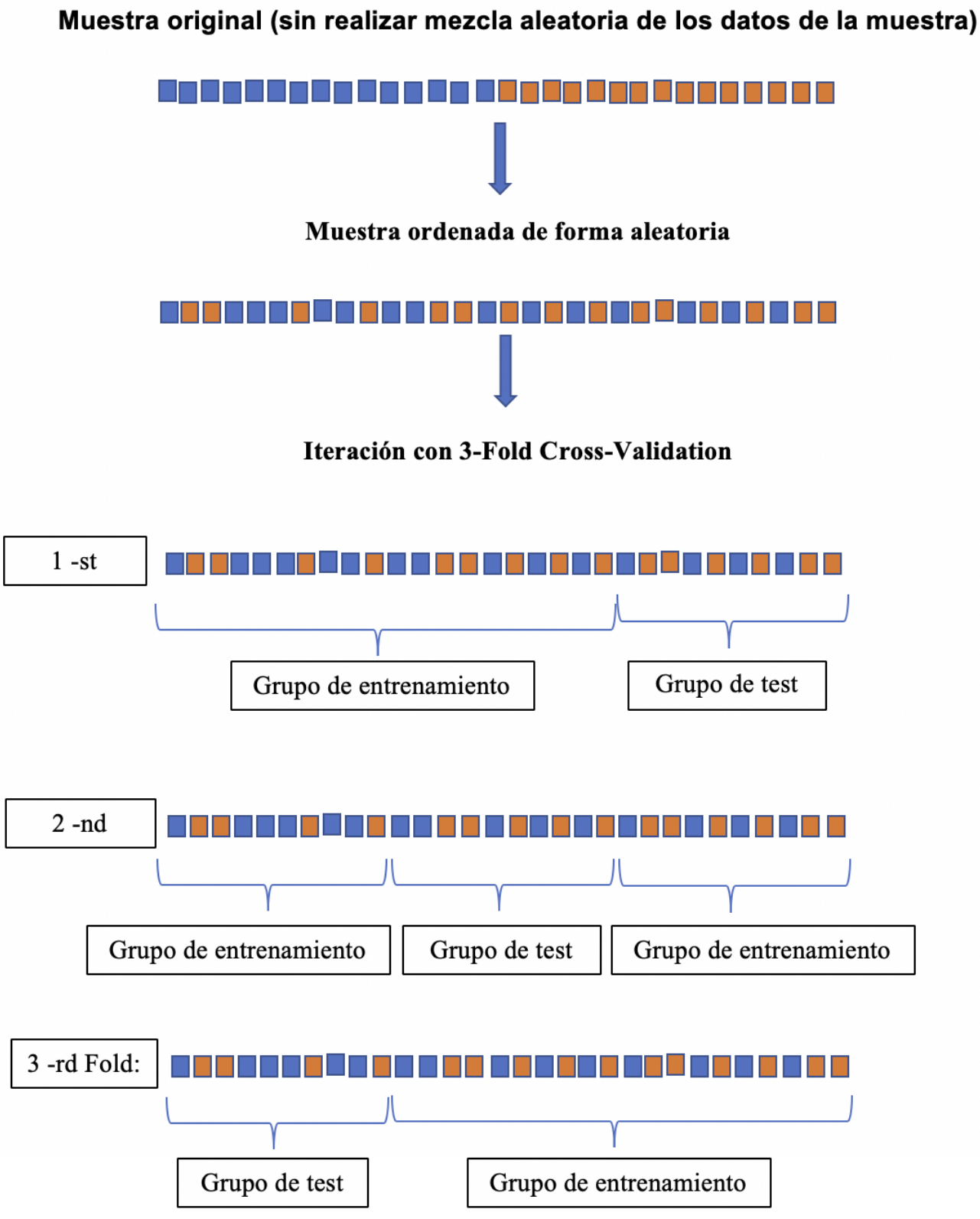


*Ilustración 25: Proceso a realizar para elaborar el modelo mediante k-Fold cross validation*

Una vez se han elaborado los 3 *Fold*, se procederá a realizar el modelo, mediante Rstudio. En este caso, el número de variables serán 6 (las cuales se comentaron anteriormente y son las variables de predicción meteorológica) y el número de nodos utilizados en esta ocasión serán 10.

El programa que se ejecutará, será similar al expuesto anteriormente, con la discrepancia que ahora se deberá realizar tres veces, cada una de ellas con su *Fold* correspondiente.

### 10.5.5 Código para el análisis mediante la herramienta Rstudio

Código utilizado para el análisis mediante Rstudio, con la finalidad de obtener con los datos obtenidos de los 3 – *Fold* los valores medios de los criterios de evaluación pertinentes:

```
# Programa para realizar la validación cruzada con 3 Fold, con los mismos datos
# que se usaron en las predicciones determinísticas


#setwd("N:/Modelos/regresion_quantregForest_3folds")

# -------

library(scoringRules)
library(readxl)
library(matrixStats)
library(stats)
library(KScorrect)
library(SpecsVerification)
library(quantreg)
library(doParallel)
library(quantregForest)
#library(xlsx)

rsq <- function (x, y) cor(x, y) ^ 2

Datos <- read_excel("3folds.xlsx")
# Datos <- read_excel("2folds.xlsx")
# Datos <- read_excel("1folds.xlsx")

df <- data.frame(Datos)
```

```
entradas_ent <- df[1:10878,6:13]
salida_ent <- df[1:10878,3]
entradas_tst <- df[10879:16317,6:13]
salida_tst <- df[10879:16317,3]

nodes <- 10 # probar con 3 a 10
mm <- 6 # probrar entre 2 y 6

qrf <- quantregForest(x=entradas_ent, y=salida_ent, nthreads=4, nodesize=nodes, mtry=mm)
conditionalQuantiles  <- predict(qrf,  entradas_ent, what = seq(0.05, 0.95, 0.05))

crps_valores <- crps_sample(salida_ent, conditionalQuantiles, method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)
crps_train <- mean(crps_valores)
prevista <- conditionalQuantiles[,10]
rmse_train <- sqrt(mean((salida_ent-prevista)^2))
mae_train <- mean(abs(salida_ent-prevista))
rsq_train <- rsq(prevista,salida_ent)

conditionalQuantiles <- predict(qrf,  entradas_tst, what = seq(0.05, 0.95, 0.05))
crps_valores <- crps_sample(salida_tst, conditionalQuantiles, method = "edf", w = NULL, bw = NULL,num_int = FALSE, show_messages = TRUE)
crps_test <- mean(crps_valores)
prevista <- conditionalQuantiles[,10]
rmse_test <- sqrt(mean((salida_tst-prevista)^2))
mae_test <- mean(abs(salida_tst-prevista))
rsq_test <- rsq(prevista,salida_tst)

rank.hist <- Rankhist(conditionalQuantiles, salida_tst)
teo <- matrix(nrow=1, ncol=20, 5339/20)
PlotRankhist(rank.hist, mode="raw")
abline(h=5339/20, col="red",lwd=3)
rmsd_test <- sqrt(1/20*sum((rank.hist-teo)^2))
```

Tras ejecutar el programa, los resultados obtenidos de los 3 - *Folds* han sido los siguientes:

| | | Fold 1 | | | | | Fold 2 | | | | | Fold 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nº nodos | Nº variables | CRPS test | RMSE test | MAE test | RSQ test | RMSD test | CRPS test | RMSE test | MAE test | RSQ test | RMSD test | CRPS test | RMSE test | MAE test | RSQ test | RMSD test |
| 3 | 2 | 51,5979 | 158,7653 | 70,9179 | 0,9159 | 29,1298 | 52,3519 | 158,3738 | 72,2363 | 0,9182 | 28,3169 | 52,0236 | 160,9981 | 70,9507 | 0,9157 | 33,2979 |
| 3 | 3 | 51,4362 | 159,5967 | 70,8662 | 0,9151 | 18,0068 | 52,2721 | 159,5443 | 72,3011 | 0,9171 | 16,7406 | 51,9989 | 161,7653 | 70,8657 | 0,9151 | 21,4111 |
| 3 | 4 | 51,4311 | 160,2869 | 70,7647 | 0,9144 | 18,5996 | 52,1492 | 159,7148 | 71,7891 | 0,9169 | 23,3783 | 52,3084 | 163,4198 | 71,5305 | 0,9134 | 23,4275 |
| 3 | 5 | 51,5303 | 160,8333 | 70,6795 | 0,9139 | 17,4484 | 52,6622 | 161,0241 | 72,3603 | 0,9156 | 16,6747 | 52,4701 | 163,8207 | 71,6441 | 0,9131 | 21,5672 |
| 3 | 6 | 51,8924 | 161,9276 | 71,0708 | 0,9128 | 18,9802 | 52,6647 | 162,0374 | 72,5054 | 0,9146 | 21,4533 | 52,6996 | 164,6761 | 72,1243 | 0,9122 | 21,9601 |
| 4 | 2 | 51,7344 | 159,4322 | 71,3504 | 0,9152 | 31,2129 | 52,3236 | 158,4136 | 72,2358 | 0,9182 | 25,4882 | 52,2432 | 161,6061 | 71,3834 | 0,9151 | 30,1835 |
| 4 | 3 | 51,2951 | 159,8795 | 70,8426 | 0,9148 | 22,9705 | 52,1357 | 159,5314 | 71,9649 | 0,9171 | 19,5051 | 52,1012 | 161,9159 | 70,9572 | 0,9149 | 23,8484 |
| 4 | 4 | 51,4769 | 160,1242 | 70,7133 | 0,9146 | 18,0706 | 52,2679 | 159,8567 | 71,9647 | 0,9167 | 19,5793 | 52,1558 | 162,4592 | 71,4205 | 0,9144 | 19,2911 |
| 4 | 5 | 51,6619 | 162,0054 | 70,9952 | 0,9127 | 24,2785 | 52,4044 | 160,6586 | 72,2209 | 0,9159 | 18,9511 | 52,5037 | 163,7638 | 71,7727 | 0,9131 | 24,4509 |
| 4 | 6 | 51,8734 | 162,4166 | 71,3248 | 0,9123 | 20,7013 | 52,6793 | 162,1795 | 72,6348 | 0,9144 | 22,7782 | 52,7822 | 164,7678 | 72,3331 | 0,9121 | 25,6875 |
| 5 | 2 | 51,7561 | 158,8183 | 71,1558 | 0,9158 | 32,1037 | 52,4206 | 158,4313 | 72,4058 | 0,9182 | 23,1289 | 52,0827 | 160,8769 | 71,0617 | 0,9159 | 31,5729 |
| 5 | 3 | 51,4746 | 159,4227 | 70,9033 | 0,9152 | 25,5685 | 52,3664 | 159,8388 | 72,2761 | 0,9167 | 18,7416 | 52,1139 | 161,5579 | 70,8943 | 0,9153 | 28,1948 |
| 5 | 4 | 51,3627 | 160,1402 | 70,7403 | 0,9146 | 23,5891 | 52,2571 | 160,0884 | 72,1071 | 0,9165 | 21,2943 | 52,1455 | 161,9907 | 71,0501 | 0,9149 | 23,8379 |
| 5 | 5 | 51,7292 | 161,6334 | 71,1029 | 0,9134 | 21,0843 | 52,4022 | 160,2929 | 72,1044 | 0,9163 | 16,9307 | 52,2541 | 163,0381 | 71,3131 | 0,9139 | 20,9248 |
| 5 | 6 | 51,8031 | 162,4735 | 71,3616 | 0,9122 | 22,1799 | 52,8109 | 161,8331 | 72,5534 | 0,9148 | 20,0287 | 52,8436 | 164,7112 | 72,0261 | 0,9122 | 19,2522 |
| 6 | 2 | 51,7468 | 158,6072 | 71,2131 | 0,9159 | 31,0974 | 52,4554 | 158,4882 | 72,3861 | 0,9181 | 26,6992 | 52,3972 | 161,5532 | 71,3539 | 0,9152 | 32,9613 |
| 6 | 3 | 51,3916 | 159,4102 | 70,9897 | 0,9153 | 22,2744 | 52,1975 | 159,0512 | 72,0814 | 0,9175 | 23,4616 | 52,2587 | 162,2438 | 71,2695 | 0,9146 | 24,3689 |
| 6 | 4 | 51,4956 | 159,9531 | 70,6641 | 0,9147 | 26,4905 | 52,3204 | 160,1215 | 72,3844 | 0,9164 | 16,7046 | 52,1239 | 162,3111 | 71,0956 | 0,9146 | 23,3377 |
| 6 | 5 | 51,6297 | 161,4612 | 71,2582 | 0,9132 | 22,0737 | 52,6335 | 161,2665 | 72,4711 | 0,9153 | 21,0914 | 52,6199 | 163,8222 | 71,8269 | 0,9131 | 27,9723 |
| 6 | 6 | 51,8724 | 162,6417 | 71,4821 | 0,9121 | 23,4363 | 52,8088 | 161,4686 | 72,6599 | 0,9151 | 20,4192 | 52,7347 | 164,6862 | 72,1264 | 0,9122 | 26,4395 |
| 7 | 2 | 51,9173 | 159,5163 | 71,6375 | 0,9149 | 29,2344 | 52,4795 | 158,1491 | 72,3548 | 0,9184 | 25,1843 | 52,5563 | 161,4595 | 71,5468 | 0,9152 | 33,9256 |
| 7 | 3 | 51,4541 | 159,5527 | 71,0054 | 0,9151 | 31,8582 | 52,3922 | 150,5075 | 72,4187 | 0,9171 | 20,1332 | 52,1867 | 162,0391 | 71,1587 | 0,9147 | 23,5042 |
| 7 | 4 | 51,5789 | 160,6019 | 71,0464 | 0,9141 | 21,0035 | 52,4771 | 160,0936 | 72,1531 | 0,9165 | 19,9261 | 52,3677 | 162,7487 | 71,4734 | 0,9141 | 21,4417 |
| 7 | 5 | 51,6987 | 161,5349 | 71,2549 | 0,9131 | 21,4859 | 52,6281 | 161,0187 | 72,8126 | 0,9155 | 15,7018 | 52,4866 | 163,1461 | 71,6544 | 0,9137 | 18,2742 |
| 7 | 6 | 51,8485 | 162,4796 | 71,5566 | 0,9122 | 19,0328 | 52,7569 | 161,4952 | 72,8121 | 0,9151 | 14,3194 | 52,6885 | 163,7558 | 71,9441 | 0,9131 | 21,6759 |
| 8 | 2 | 51,9118 | 159,0079 | 71,6666 | 0,9155 | 34,6331 | 52,8606 | 159,4222 | 72,9585 | 0,9171 | 28,5701 | 52,5944 | 161,4574 | 71,4503 | 0,9152 | 31,1603 |
| 8 | 3 | 51,6512 | 160,2906 | 71,2688 | 0,9143 | 23,0271 | 52,4696 | 159,2362 | 72,5431 | 0,9173 | 23,0899 | 52,2153 | 161,5202 | 71,0342 | 0,9153 | 31,1969 |
| 8 | 4 | 51,5306 | 160,2507 | 71,0145 | 0,9144 | 23,9488 | 52,3786 | 160,0797 | 72,4265 | 0,9165 | 21,1862 | 52,3903 | 162,7443 | 71,6511 | 0,9141 | 24,6363 |
| 8 | 5 | 51,5822 | 161,1201 | 71,1108 | 0,9135 | 20,8242 | 52,6852 | 160,7761 | 72,7205 | 0,9158 | 15,6316 | 52,5511 | 163,0919 | 71,6193 | 0,9137 | 27,7101 |
| 8 | 6 | 51,9256 | 162,8296 | 71,8058 | 0,9118 | 18,3588 | 52,8397 | 162,0974 | 72,9838 | 0,9144 | 21,4929 | 52,8202 | 164,6457 | 72,3031 | 0,9122 | 17,2031 |
| 9 | 2 | 52,0803 | 160,0612 | 71,8645 | 0,9144 | 31,9742 | 52,7126 | 158,8908 | 72,9951 | 0,9177 | 29,7934 | 52,4898 | 160,9477 | 71,3669 | 0,9157 | 39,6541 |
| 9 | 3 | 51,6173 | 159,7914 | 71,3629 | 0,9148 | 27,9473 | 52,4573 | 159,6523 | 72,6709 | 0,9169 | 18,2498 | 52,2535 | 161,1735 | 71,1796 | 0,9156 | 28,3963 |
| 9 | 4 | 51,7052 | 161,1444 | 71,4266 | 0,9135 | 25,0907 | 52,5243 | 160,1839 | 72,4489 | 0,9164 | 19,8405 | 52,4061 | 162,6201 | 71,5085 | 0,9142 | 26,7871 |
| 9 | 5 | 51,7501 | 161,9768 | 71,5788 | 0,9126 | 25,6427 | 52,6726 | 160,7996 | 72,7368 | 0,9158 | 22,8877 | 52,6253 | 163,7673 | 72,0143 | 0,9131 | 22,2856 |
| 9 | 6 | 52,0022 | 162,9078 | 71,7652 | 0,9117 | 18,9432 | 52,9402 | 161,9542 | 72,9428 | 0,9164 | 21,4417 | 52,9004 | 164,4982 | 72,3793 | 0,9123 | 21,7749 |
| 10 | 2 | 52,1335 | 159,7594 | 72,3319 | 0,9148 | 32,1721 | 52,7046 | 158,9518 | 72,9557 | 0,9176 | 28,1682 | 52,5337 | 161,0727 | 71,5513 | 0,9156 | 33,5119 |
| 10 | 3 | 51,6451 | 159,9194 | 71,2888 | 0,9146 | 25,5489 | 52,4417 | 158,9407 | 72,4064 | 0,9176 | 23,6039 | 52,4163 | 162,1344 | 71,4017 | 0,9146 | 27,9275 |
| 10 | 4 | 51,7855 | 160,9974 | 71,3969 | 0,9136 | 25,2952 | 52,6161 | 160,2297 | 72,6004 | 0,9163 | 18,8667 | 52,3329 | 162,7845 | 71,3525 | 0,9141 | 23,5594 |
| 10 | 5 | 51,8377 | 161,7492 | 71,3902 | 0,9129 | 27,9883 | 52,7369 | 160,9925 | 72,8855 | 0,9155 | 22,3841 | 52,5633 | 163,3905 | 71,8096 | 0,9134 | 27,6794 |
| 10 | 6 | 51,8897 | 162,0894 | 71,6148 | 0,9126 | 23,9843 | 52,7985 | 161,6041 | 72,9751 | 0,9149 | 22,8505 | 52,8559 | 164,4735 | 72,3409 | 0,9123 | 19,8783 |

***Tabla 14: Resultados obtenidos de los 3-Folds a través de Rstudio***

Después de obtener los resultados de los 3 – *Folds*, se realiza la media de cada uno de los criterios evaluadores obtenidos para cada nodo con su combinación de diferentes números de variables; para así conseguir un criterio que pueda ser evaluado para determinar que combinación de nodo y número de variables es la que mejores predicciones originará para aplicar posteriormente sobre los datos del grupo de test. Los valores medios obtenidos de cada uno de los criterios evaluadores se representarán en la siguiente tabla:

| | | Valores medios | | | | |
|---|---|---|---|---|---|---|
| Nº nodos | Nº variables | CRPS medio | RMSE medio | MAE medio | RSQ medio | RMSD medio |
| 3 | 2 | 51,9911 | 159,3791 | 71,3683 | 0,9166 | 30,2482 |
| 3 | 3 | 51,9024 | 160,3021 | 71,3443 | 0,9158 | 18,7195 |
| 3 | 4 | 51,9629 | 161,1405 | 71,3614 | 0,9149 | 21,8018 |
| 3 | 5 | 52,2209 | 161,8927 | 71,5613 | 0,9142 | 18,5634 |
| 3 | 6 | 52,4189 | 162,8804 | 71,9002 | 0,9132 | 20,7979 |
| 4 | 2 | 52,1004 | 159,8173 | 71,6565 | 0,9162 | 28,9615 |
| 4 | 3 | 51,8440 | 160,4423 | 71,2549 | 0,9156 | 22,1080 |
| 4 | 4 | 51,9669 | 160,8134 | 71,3662 | 0,9152 | 18,9803 |
| 4 | 5 | 52,1900 | 162,1426 | 71,6629 | 0,9139 | 22,5602 |
| 4 | 6 | 52,4450 | 163,1213 | 72,0976 | 0,9129 | 23,0557 |
| 5 | 2 | 52,0865 | 159,3755 | 71,5411 | 0,9166 | 28,9352 |
| 5 | 3 | 51,9850 | 160,2731 | 71,3579 | 0,9157 | 24,1683 |
| 5 | 4 | 51,9218 | 160,7398 | 71,2992 | 0,9153 | 22,9071 |
| 5 | 5 | 52,1285 | 161,6548 | 71,5068 | 0,9145 | 19,6466 |
| 5 | 6 | 52,4859 | 163,0059 | 71,9804 | 0,9131 | 20,4869 |
| 6 | 2 | 52,1998 | 159,5495 | 71,6510 | 0,9164 | 30,2526 |
| 6 | 3 | 51,9493 | 160,2351 | 71,4469 | 0,9158 | 23,3683 |
| 6 | 4 | 51,9800 | 160,7952 | 71,3814 | 0,9152 | 22,1776 |
| 6 | 5 | 52,2944 | 162,1833 | 71,8521 | 0,9139 | 23,7125 |
| 6 | 6 | 52,4720 | 162,9322 | 72,0895 | 0,9131 | 23,4317 |
| 7 | 2 | 52,3177 | 159,7083 | 71,8464 | 0,9162 | 29,4481 |
| 7 | 3 | 52,0110 | 157,3664 | 71,5276 | 0,9156 | 25,1652 |
| 7 | 4 | 52,1412 | 161,1481 | 71,5576 | 0,9149 | 20,7904 |
| 7 | 5 | 52,2711 | 161,8999 | 71,9073 | 0,9141 | 18,4873 |
| 7 | 6 | 52,4313 | 162,5769 | 72,1043 | 0,9135 | 18,3427 |
| 8 | 2 | 52,4556 | 159,9625 | 72,0251 | 0,9159 | 31,4545 |
| 8 | 3 | 52,1120 | 160,3490 | 71,6154 | 0,9156 | 25,7713 |
| 8 | 4 | 52,0998 | 161,0249 | 71,6974 | 0,9150 | 23,2571 |
| 8 | 5 | 52,2728 | 161,6627 | 71,8169 | 0,9143 | 21,3886 |
| 8 | 6 | 52,5285 | 163,1909 | 72,3642 | 0,9128 | 19,0183 |
| 9 | 2 | 52,4276 | 159,9666 | 72,0755 | 0,9159 | 33,8072 |
| 9 | 3 | 52,1094 | 160,2057 | 71,7378 | 0,9158 | 24,8645 |
| 9 | 4 | 52,2119 | 161,3161 | 71,7947 | 0,9147 | 23,9061 |
| 9 | 5 | 52,3493 | 162,1812 | 72,1100 | 0,9138 | 23,6053 |
| 9 | 6 | 52,6143 | 163,1201 | 72,3624 | 0,9135 | 20,7199 |
| 10 | 2 | 52,4573 | 159,9280 | 72,2796 | 0,9160 | 31,2841 |
| 10 | 3 | 52,1677 | 160,3315 | 71,6990 | 0,9156 | 25,6934 |
| 10 | 4 | 52,2448 | 161,3372 | 71,7833 | 0,9147 | 22,5738 |
| 10 | 5 | 52,3793 | 162,0441 | 72,0284 | 0,9139 | 26,0173 |
| 10 | 6 | 52,5147 | 162,7223 | 72,3103 | 0,9133 | 22,2377 |

*Tabla 15: Valores medios obtenidos de los 3-Folds, determinación de la combinación de nodos y variables óptima*

Después de inspeccionar los resultados obtenidos y, tendiendo en cuenta los criterios de evaluación determinados para los modelos de predicción, se ha decidido que la combinación de variables y nodos que ha proporcionado una solución más óptima, ha sido la de 4 nodos y 3 variables, cuyos resultados después de aplicar dicha combinación con los datos de la muestra del grupo de test, son los siguientes:

| Nº nodos | Nº variables | CRPS entren | RMSE entren | MAE entren | $R^2$ entren | CRPS test | RMSE test | MAE test | $R^2$ test | RMSD test |
|---|---|---|---|---|---|---|---|---|---|---|
| **4** | **3** | 10,8676 | 21,9193 | 2,9324 | 0,9984 | 52,6431 | 151,5386 | 71,0676 | 0,9348 | 54,4264 |

*Tabla 16: Resultados finales de predicción del grupo de entrenamiento y test, modelo Random Forest*

# 11. Comparación de los modelos de predicción probabilístos

Los modelos que se han elaborado en este trabajo final de grado, no son extrapolables a cualquier planta fotovoltaica o datos puedan tenerse en relación a potencia producida, sino que han sido creados para determinar la generación de la planta de Alcolea del Río.

Otro punto a tener en cuenta, es que los modelos no son generalizables para cualquier periodo de tiempo, pues, los modelos que han sido expuestos en este trabajo han sido elaborados entre el 1 de julio de 2016 y el 9 de marzo de 2019.

En primer lugar, se realizará la comparación entre los modelos cuya predicción se ha elaborada para predecir los valores del día D. Dentro de este grupo, se encuentra el modelo climatológico, el cual permite obtener los resultados del cualquier día del año a cualquier hora, independiente de cualquier variable meteorológica (este modelo se incluye en la comparación de ambos grupos) y el modelo persistente probabilístico de día similar “hoy-hoy”.

Dichos resultados se compararán en términos de error absoluto medio (MAE), de la raíz de la desviación cuadrática media (RMSE), del coeficiente de determinación ($R^2$), la puntuación de probabilidad de clasificación continua (CRPS) y del RMSD, se presentarán en la tabla siguiente:

| | | CRPS | RMSE | MAE | $R^2$ | RMSD |
|---|---|---|---|---|---|---|
| **Modelo climatológico** | **Entrenamiento** | 83,19345 | 248,83940 | 116,9331 | 0,8185 | 23,4391 |
| | **Test** | 60,64788 | 164,24777 | 81,0676 | 0,9169 | 51,5445 |
| **Modelo persistente probabilístico "hoy - hoy"** | **Entrenamiento** | 160,2174 | 210,9737 | 83,8204 | 0,8602 | 127,6114 |
| | **Test** | 45,54270 | 154,10031 | 59,4432 | 0,92837 | 28,39084 |

*Tabla 17: Comparación de los modelos para el día D (predicciones para hoy)*

Para facilitar la interpretación y comparativa de los modelos estudiados se han representado los resultados de éstos en el siguiente gráfico:

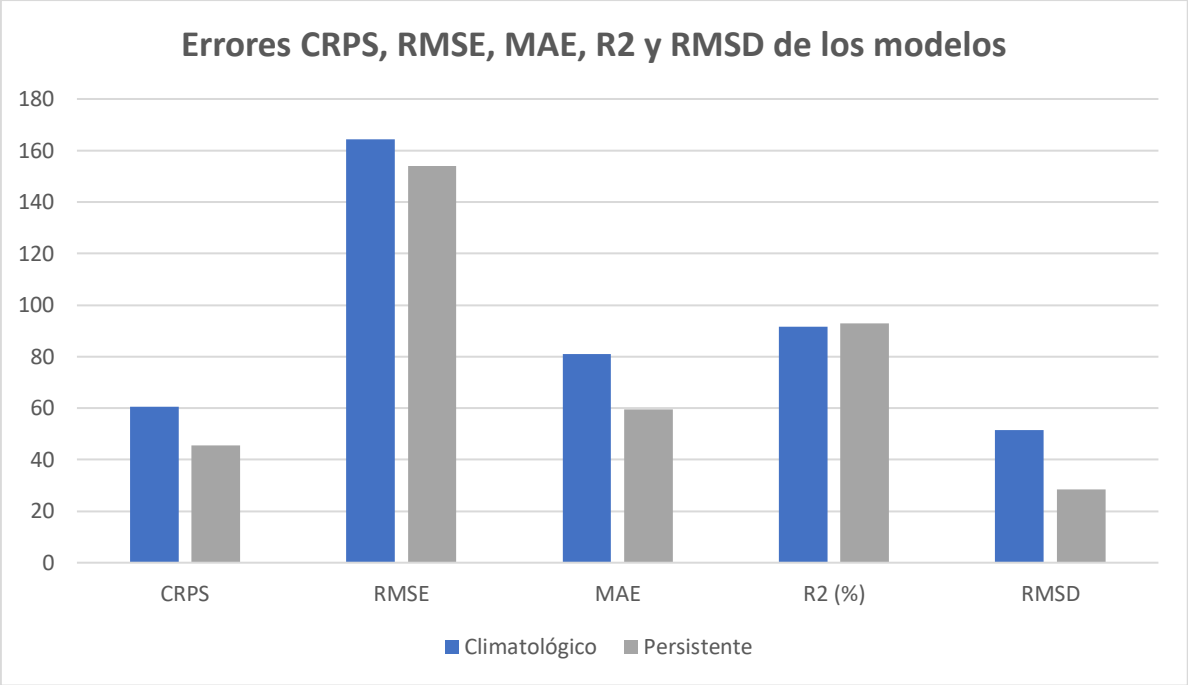


*Ilustración 26: Representación de los errores para los modelos de predicción del día D*

Después de analizar los datos obtenidos tras evaluar los modelos con los datos del grupo de test, teniendo en cuenta los criterios evaluadores que se han explicado a lo largo de este trabajo y recalcando el compromiso de obtener el mínimo CRPS y RMSD, se deduce que entre el modelo climatológico y el modelo persistente probabilístico el que mejores resultados ofrece es el modelo persistente probabilístico.

Posiblemente la mejora en estos criterios por parte del modelo persistente probabilístico reside en la utilización de la predicción de las variables meteorológicas junto con los datos del grupo de entrenamiento de la planta fotovoltaica ofreciendo una mayor información que facilita el realizar una predicción más fiable y precisa.

Por otro lado, el modelo climatológico es relativamente sencillo y tiene ciertas limitaciones, ya que este método ofrece una técnica simple para predecir el clima porque se basa en tendencias pasadas y predicen el clima para un día y ubicación específicos en función de las condiciones climáticas para ese mismo día durante varios años en el pasado, es decir, este modelo con los datos con los que se han trabajado siempre va a determinar el mismo valor, independiente de la hora, día, año que sea.

Ahora, se efectuará la comparación entre los modelos cuya predicción se ha elaborada para predecir los valores del día D+1, es decir, la predicción de los valores para mañana. Dentro de este grupo, se encuentra el modelo climatológico, el cual permite obtener los resultados del cualquier día del año a cualquier hora, independiente de cualquier variable meteorológica (este modelo se incluye en la comparación de ambos grupos), el modelo persistente probabilístico de día similar “hoy-mañana”, la regresión de cuantiles, la regresión lineal múltiple y el Random Forest.

Dichos resultados se compararán en términos de error absoluto medio (MAE), de la raíz de la desviación cuadrática media (RMSE), del coeficiente de determinación ($R^2$), la puntuación de probabilidad de clasificación continua (CRPS) y del RMSD, se presentarán en la tabla siguiente:

| | | CRPS | RMSE | MAE | $R^2$ | RMSD |
|---|---|---|---|---|---|---|
| **Modelo climatológico** | **Entrenamiento** | 83,19345 | 248,83940 | 116,9331 | 0,8185 | 23,4391 |
| | **Test** | 60,64788 | 164,24777 | 81,0676 | 0,9169 | 51,5445 |
| **Modelo persistente probabilístico “hoy - mañana”** | **Entrenamiento** | 215,8542 | 219,4222 | 88,6613 | 0,8522 | 202,2411 |
| | **Test** | 45,54270 | 154,10031 | 59,4432 | 0,92837 | 28,39084 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Regresión de cuantiles** | **Entrenamiento** | 86,0312 | 222,6523 | 113,7515 | 0,8423 | - |
| | **Test** | 79,0231 | 181,0470 | 98,7776 | 0,9189 | - |
| **Regresión lineal múltiple** | **Entrenamiento** | 102,1798 | 218,5122 | 130,5623 | 0,8437 | - |
| | **Test** | 93,8734 | 200,4418 | 125,3817 | 0,9179 | - |
| **Random Forest** | **Entrenamiento** | 10,8676 | 21,9193 | 2,9324 | 0,9984 | - |
| | **Test** | 52,6431 | 151,5386 | 71,0676 | 0,9348 | 54,4264 |

*Tabla 18: Comparación de los resultados finales entre los modelos de predicción para el día D+1*

Para facilitar la interpretación y comparativa de los modelos estudiados se han representado los resultados de éstos en el siguiente gráfico:

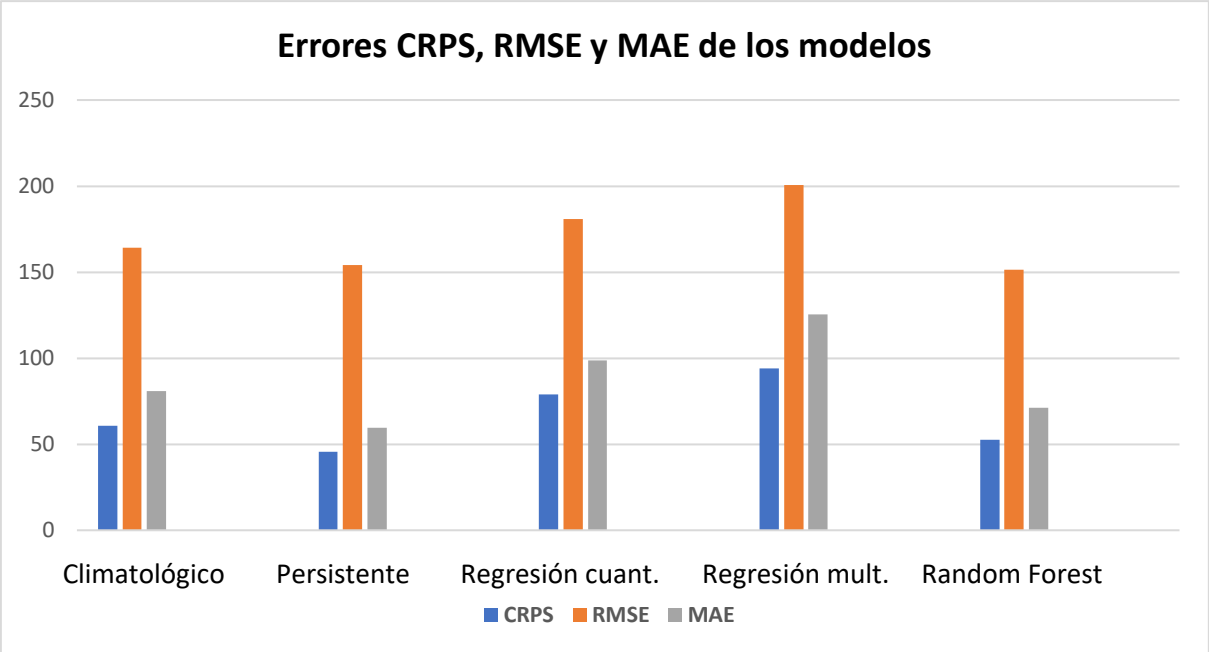


*Ilustración 27: Representación de los errores de los modelos de predicción para el día D+1*

Después de analizar los datos obtenidos tras evaluar los modelos con los datos del grupo de test, teniendo en cuenta los criterios evaluadores que se han explicado a lo largo de este trabajo y recalcando el compromiso de obtener el mínimo CRPS, MAE y RMSD, junto a maximizar el $R^2$, se concluye que hay tres modelos que obtienen unos resultados de predicción bastante buenos. Estos

modelos son el modelo climatológico, el modelo persistente probabilístico “hoy-mañana” y el Random Forest y, tras revisar los datos se deduce que el que mejores resultados ofrece es el modelo persistente probabilístico “hoy-mañana”.

Es curioso como este modelo ha resultado ser el que mejores predicciones realiza tanto en predicción de valores para el día D, como la predicción de datos para el día D+1. Además, obtiene mejores resultados que el modelo climatológico, algo que no es raro, puesto que este es el modelo más sencillo, el cuál siempre ofrece los mismos resultados para las mismas horas. Sin embargo, resulta interesante que obtenga mejores resultados a la hora de realizar predicciones frente a un modelo más complejo y sofisticado como es el Random forest.

# 12. Conclusiones

Se han desarrollado un conjunto de modelos para la obtención de predicciones probabilísticas para una planta fotovoltaica. Estos modelos incluyen desde los desarrollados con sencillas técnicas estadísticas (climatológico, persistente o regresión de cuantiles) hasta los más modernos y sofisticados (random forest).

Los modelos se han aplicado a los datos reales de producción de una planta fotovoltaica situada en Alcolea del Río (Sevilla). Los datos se dividieron en dos grupos: los datos de ajuste o entrenamiento y los datos de prueba o test. Estos últimos no se han utilizado para entrenar o ajustar los modelos, por lo que la comparación de los resultados de los modelos sobre estos datos de test permite una comparación fidedigna del funcionamiento de los mismos y permite establecer cual de los modelos tiene un mejor comportamiento predictivo.

Como ya se comentó anteriormente, hay dos grupos entre los modelos de predicción, los que realizan la predicción para el día D y los que realizan la predicción para el día D+1. Dentro del primer grupo, solo se encuentran el modelo climatológico y el modelo persistente probabilístico de “Hoy-hoy”. Entre ellos, el que mejores resultados predictivos ha obtenido ha sido el modelo persistente probabilístico.

Cabe destacar que, en ninguno de los modelos se ha detectado un sobre entrenamiento (overfitting) que podría haber perjudicado a los resultados cuando

han sido evaluados por los datos del grupo de test y tampoco ha quedado reflejado en ninguno de los criterios evaluadores. Destacar que, entre ambos modelos las diferencias existentes entre los criterios evaluadores (CRPS, RMSE, MAE…) no son excesivamente grandes, pero uno de estos criterios si que denota un poco la diferencia y es el RMSD, dicho criterio junto con el CRPS son los que proporcionan precisión, fiabilidad y nitidez a las predicciones realizadas.

Asimismo, destacar que el modelo climatológico al ser el más sencillo de todos refleja unas limitaciones, el cual no utiliza ninguna de las predicciones meteorológicas lo que restringe su capacidad para poder ajustarse mejor a los datos reales.

El otro grupo que elabora las predicciones para el día D+1, está formado por el modelo climatológico, el modelo persistente probabilístico de “Hoy-mañana”, el modelo de regresión de cuantiles, el modelo de regresión lineal múltiple y el random forest. En cuanto a los resultados obtenidos, es curioso ver como es otra vez el modelo persistente probabilístico el que mejores resultados predictivos muestra, lo que determina en cierto modo que no por tratarse de un modelo más sofisticado y complejo, como el random forest, se deben obtener mejores resultados.

Respecto a los resultados obtenidos a través de los criterios evaluadores, reseñar que el modelo random forest ofrece cierto sobre entrenamiento (overfitting), que en cierto modo ha podido perjudicar con tales efectos a los resultados predictivos obtenidos al evaluar el modelo con los datos del grupo de test. Los modelos de regresión son los que peores resultados obtuvieron, tanto con los datos de entrenamiento como con los de test, al igual que el modelo climatológico, estos modelos tienen sus limitaciones y llegado un momento, aunque se introdujeran más variables o parámetros para realizar las predicciones los resultados no sufrirían cambios significativos.

Por tanto, el modelo que ha proporcionado los mejores resultados en ambos grupos ha sido el persistente probabilístico, la razón más probable para que este modelo haya sido mejor incluso que el random forest es la estabilidad atmosférica en el lugar donde esta la planta (Alcolea del Río) de un día para otro. En Sevilla se suceden los días soleados de forma frecuente (si la planta estuviese en Galicia saldría un resultado muy diferente). Por ello, al parecerse mucho un día a los anteriores (a efectos de la temperatura y radiación solar principalmente) parece lógico que este modelo sea el que mejores resultados ofrezca.

Asimismo, destacar la importancia de la aparición de este tipo de modelos de predicción y de su elección para la realización de este trabajo final de grado,

porque proporcionan una información completa que puede ser de gran utilidad para evaluar el riesgo económico en operaciones en el mercado (venta de la energía generada), permitiendo analizar de antemano las probabilidades de obtener o no los beneficios perseguidos (por el tema de las penalizaciones en el mercado eléctrico por la no generación de los valores de energía ofertados).

# 13. Bibliografía y webgrafía

## 13.1 Listado de la bibliografía

Libros:


- Bassett, G., Koenker, R. An empirical quantile function for linear models with iid errors. Journal of the American Statistical Association 77(378), pp. 407-415, 1982.

- Buchinsky, M. Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963-1987. Journal of Econometrics 65(1), pp. 109-154, 1995.

- Graham Williams. Data Mining with Rattle and R. Springer, 2011.

- Tsau Young Lin, Ying Xie, Anita Wasilewska y Churn-Jung Liau. Data Mining: Foundations and Practice. Springer, 2008.

- Robert Nisbet, John Elder y Gary Miner. Handbook of statistical analysis and data mining applications. Elsevier, 2009.

- Ning Zhang, Chongqing Kang, Ershun Du y Yi Wang. Analytics and optimization for Renewable Energy Integration. CRC Press, 2019.
- Genuer R., Poggi JM. Tuleau-Malot C."Variable Selection Using Random Forests"(2010). Bourdeaux University.

## 13.2 Listado de la webgrafía

Artículos de interés:


- Crecimiento de la energía solar fotovoltaica instalada: https://www.energias-renovables.com/fotovoltaica/crece-en-dos-anos-casi-un-500-20190204

- Información sobre Vallehermoso I, planta fotovoltaica de Alcolea del Río: https://drive.google.com/file/d/0B3bp-NX99raBeElvRDkzTnhCb0RBM0Qzc2doaklqQThmWUpF/view

- Previsión meteorológica en la energía solar fotovoltaica: http://www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/publicaciones/Fisica_del_caos_en_la_predicc_meteo/38_Aplicaciones_en_energia_solar.pdf

- Validación de una herramienta para la predicción de energía solar en horizontes de tiempo cercanos: https://www.smartgridsinfo.es/comunicaciones/comunicacion-validacion-herramienta-prediccion-energia-solar-horizontes-tiempo-cercanos

- Cuatro tipos de pronóstico: https://sciencing.com/four-types-forecasting-8155139.html

- Probabilistic forecasting of day-ahead solar irradiance using quantile gradient boosting: https://pdfs.semanticscholar.org/b3b7/70c237c27152e405dcf6f4063f91b7712f25.pdf

2018-2019